CULTURAL IMPLICATIONS OF SELF-OTHER AGREEMENT IN MULTISOURCE FEEDBACK: COMPARING SAMPLES FROM U.S., CHINA, AND GLOBALLY DISPERSED TEAMS

Yue Lin, B.S., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2007

APPROVED:

Michael M. Beyerlein, Major Professor
Bert Hayslip, Jr. Committee Member
Craig S. Neumann, Committee Member
Timothy P. Carey, Committee Member
Linda Marshall, Chair of the Department of Psychology
Sandra L. Terrell, Dean of the Robert B. Toulouse School of Graduate Studies

Lin, Yue. Cultural implications of self-other agreement in multisource feedback: Comparing samples from US, China, and globally dispersed teams. Doctor of Philosophy (Industrial/Organizational Psychology), August 2007, 190 pp., 15 tables, 4 illustrations, references, 235 titles.

Application of multisource feedback (MSF) increased dramatically and became widespread globally in the past two decades, but there was little conceptual work regarding self-other agreement and few empirical studies investigated self-other agreement in other cultural settings. This study developed a new conceptual framework of self-other agreement and used three samples to illustrate how national culture affected self-other agreement. These three samples included 428 participants from China, 818 participants from the US, and 871 participants from globally dispersed teams (GDTs). An EQS procedure and a polynomial regression procedure were used to examine whether the covariance matrices were equal across samples and whether the relationships between self-other agreement and performance would be different across cultures, respectively. The results indicated MSF could be applied to China and GDTs, but the pattern of relationships between self-other agreement and performance was different across samples, suggesting that the results found in the U.S. sample were the exception rather than rule. Demographics also affected self-other agreement disparately across perspectives and cultures, indicating self-concept was susceptible to cultural influences. The proposed framework only received partial support but showed great promise to guide future studies. This study contributed to the literature by: (a) developing a new framework of self-other agreement that could be used to study various contextual factors; (b) examining the relationship between self-other agreement

and performance in three vastly different samples; (c) providing some important insights about consensus between raters and self-other agreement; (d) offering some practical guidelines regarding how to apply MSF to other cultures more effectively.

ACKNOWLEDGMENTS

There are many people who have contributed to the successful completion of this dissertation. I would like to express my deepest gratitude to my committee chairperson, Dr. Michael Beyerlein for his knowledge, guidance, encouragement, and commitment to the student. I would also like to thank my committee members for taking the time to contribute their thoughts to this work. Every committee member was very responsive and accommodative. As a special note, Dr. Carey got up early morning to attend the meetings from Hong Kong, and he also contributed to correcting some inconsistencies and editorial errors in the draft. I am grateful to Personnel Decisions International Corporation who agreed to share the data for this research. Special thanks also go to Paul Van Katwyk and Helen Chen for their commitment to employee's personal growth. I would also like to thank my former college classmates, Wanjiang Du and Guorong Zhu, for giving me generous support and advice.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

CHAPTER I

INTRODUCTION

Feedback is critical to the survival of all systems, regardless of whether they are individuals or institutions. A system is composed of four basic components: input, throughput, output, and feedback loop (Katz & Kahn, 1978). The feedback loop runs from output to input to control variations and keep the system in equilibrium. The breakdown of the feedback loop would cause the system's dissipation. For individuals, the feedback loop is innately built in and can operate automatically. For institutions, however, the feedback loop has to be designed and built in with intention. Performance appraisal is intended to serve such a feedback function.

As success of an organization depends on efforts from each individual that is composed of the organization, the organization needs to make sure everyone performs to the expectation. An intuitive approach is to measure the employee's output quality and quantity, a method can be traced back to time study (Taylor, 1911/1947) and motion study (Gilbreth, 1909, 1911). However, objective evidence does not always exist (Festinger, 1954). Objective outcome measures for individuals become even harder to define when interdependency among employees increases as tasks become more complex. As a result, indirect, or more subjective measures, become surrogates for objective outcome measures. To ensure the accuracy of subjective measures, a standardized procedure,

named performance appraisal, is adopted by the organization to evaluate, judge, or estimate each employee's degree of meeting expectations and goals. Performance appraisal has been adopted widely by organizations (Teel, 1980).

However, dissatisfaction with traditional performance appraisal increased in recent years (Coens & Jenkins, 2000). Some researchers (e.g., Deming, 1986) insist that the use of performance appraisal should discontinue. There are several major concerns. The first one is about the accuracy of performance appraisal. Managers tend to assign uniform ratings to employees regardless of performance (Murphy, 1992). The second concern is related to the changing nature of the job. As tasks become more complex and interdependency among employees gets higher, considering only the manager's perspective appears quite narrow. As the scope of performance expands (Borman & Motowidlo, 1993), managers face more challenges to evaluate their subordinates' performance. Third, performance appraisal is a top-down process that served solely organizational purposes (Hedge, Borman & Birkeland, 2001). The results of appraisal are tied to compensation, promotion, or even negative employment decisions. Employees feel threatened and stigmatized when they have to sit with supervisors to discuss their performance (Meyer, 1991). Antonioni (1994) pointed out that almost one third of managers believed their performance appraisals were very ineffective and more than one third of companies had stopped using or are considering *abolishing* performance appraisal.

Various approaches have been adopted to mitigate the aforementioned problems. First, more format forms are created to reduce rating errors (Austin & Villanova, 1992). Although formats are helpful in controlling bias, this approach is too narrowly focused (Feldman, 1981; Landy & Farr, 1980). Second, different perspectives of performance are incorporated in performance ratings (Landy & Farr, 1980). Bernardin, Dahmus, and Redmon (1993) observed that managers react more favorably to ratings from multiple sources than from only one source. Third, performance appraisal is designed to serve different purposes, which generally can be categorized into developmental and administrative (Campbell, 2001). Giving development-oriented feedback based on the performance appraisal data can make performance appraisal meaningful to both the organization and the employee. Multisource Feedback (MSF), which incorporates multiple perspectives and serves mainly developmental purposes, shows a great promise in overcoming the weaknesses of traditional performance appraisal.

MSF is also called 360 degree feedback, multirater assessment, multirater feedback, multisource assessment, full-circle feedback, upward feedback, etc. (Bracken, Timmreck & Church, 2001a). MSF refers to “the practice of involving multiple raters, often including self rating, in the assessment of individuals” (Tornow, 1993a, p. 211). Rater could include supervisors, peers, subordinates, customers, and others who are familiar with the ratee’s performance and whose perspectives are valuable to the organization and the individual, because validity of ratings depends on the opportunity to observe the individual’s work through

interaction on the job. All raters should use a standardized assessment instrument to rate the performance of the ratee. MSF has received more acceptance from users (Maurer & Tarulli, 1996), and seems more legally defensible than traditional performance appraisal (Brown, n.d.).

Antonioni (1996) reported that an estimated 25% of companies were using some type of upward or 360-degree-feedback process. According to Yammarino and Atwater (1997), a majority of Fortune 500 companies employed 360-degree feedback. They listed more than 50 of the world's most prestigious companies who had adopted MSF, such as, IBM, 3M, Johnson & Johnson, and Motorola. The best companies choose to develop their own executive talent inside and MSF is seen as one of the most effective development tools along with mentoring and coaching (Handfield-Jones, 2000). A recent estimation shows that over one third of U.S. companies are using some type of MSF process for managers (Bracken, Timmreck & Church, 2001a). Application of MSF has also escalated as multinational companies (MNCs) take the practice to the globe (Gillespie, 2005).

With increased application in the field, research on MSF has grown dramatically. The research momentum is evident in special issues of journals and edited books, such as *Human Resource Management* around 360 degree feedback (Tornow 1993b), *Group and Organizational Management* (Church & Bracken, 1997), and *The Handbook of Multisource Feedback: The Comprehensive Resource for Designing and Implementing MSF Processes* (Bracken, Timmreck, & Church, 2001b). Articles in the special issues and the

handbook represent the scope and depth of current research on MSF. Topics go beyond rating accuracy to include those that concern practitioners the most in the field, such as self-other agreement, measurement equivalence, improvements after feedback, MSF application in organizational change. Studies based on field data have increased dramatically since 1990s, with more cooperative efforts between researchers and practitioners.

Introducing multiple perspectives to performance appraisal clearly is an advantage of MSF, but it has some liabilities as well. The most prominent one is how to interpret rating differences between sources, particularly, self-other agreement or disagreement. A natural tendency is to see lack of agreement as errors. This view can be traced back to Cronbach's criticisms of accuracy research in person perception (Kenny & Albright, 1987). MSF researchers (e.g., Yammarino & Atwater, 1993; Church, 1994; Church, 1997; Fletcher, 1997; Fletcher & Baldry, 2000; Wohlers & London, 1989) argue that lack of agreement or discrepancy between ratings is not error but something meaningful. For example, self-other agreement could be an index of self-awareness (Wohlers &London, 1989), where self-awareness is defined as the extent of individual's understanding of his/her own strengths and weaknesses. Seeing self-other agreement as something meaningful instead of error has provoked strong research interest in the MSF field, but little work exists regarding the self-other agreement process. Several models have been proposed to clarify these ambiguities (e.g., Atwater & Yammarino, 1997; Campbell, 1985, cited in

Campbell, 1988; Yammarino & Atwater, 1993), but they generally fail to paint a clear picture how self-other agreement is achieved and what may affect the process. More theory development and empirical research are expected.

The issue of self-other agreement gets even more complicated when MSF reaches to a global scale. Applying MSF to other cultures may face many challenges. For example, is measurement equivalent across cultures? Are leadership styles generalizable in different cultures? Do globally dispersed teams have the same understanding of MSF? What are the cultural implications for self-other agreement? These questions are meaningful to explore. Fletcher (1997) summarized some cultural studies in the area of performance appraisal. He lamented that very little empirical work regarding cultural effects on performance appraisal existed. While there were some studies being conducted between western countries, few were done in non-western countries. Cross-cultural research on self-other agreement was even more rare. Levy and Williams (2004) identified numerous contextual variables that could affect performance appraisal but located few cross-cultural studies. A more recent example is a study conducted by Atwater et al. (2005), who explored the relationship between self-other agreement and performance in the U.S. and European nations. Atwater et al. (2005) also noted that very little research on self-other agreement has been conducted outside the U.S.

This study developed a conceptual framework that took culture's impact on self-other agreement into consideration and applied empirical data to validate

the framework. This effort is aligned with Landy and Farr's (1980) and Feldman's (1981) suggestion that more process and context research are needed in performance appraisal. This study can not only deepen our understanding of self-other agreement but also provide some practical guidance to applications of MSF in other cultures.

CHAPTER II

LITERATURE REVIEW

Historical Roots and Current Forces

Performance appraisal in the modern sense doesn't have a long history but its antecedents have existed for a long time. To have a clear understanding of MSF, it's important to know the pieces that compose MSF. The first segment will introduce the historical roots of MSF and then describe some current forces that shape MSF to date. A phrase of *evaluation system* is used to refer to a more ancient form of performance appraisal, in order to distinguish it from the more modern one, which is the focus of this study.

Historical Roots of MSF

The earliest evaluation system could date back to thousands of years ago in China. Various criteria and institutions were established to evaluate bureaucrats' performance in different historic periods. The evaluation system provided means and standards to identify high and poor performers, who would then be rewarded or punished based on the evaluation results. Such a practice lasted thousands of years and is still in use in modern Chinese government. This method was introduced to civil service in Europe and U.S. in late 1800s. Around 1900, the Federal Civil Service of the United States began giving merit ratings to its employees and set up special institutions to manage it (Wiese & Buckley,

1998). Meanwhile, psychologists started playing role in this fledging field through assisting the military to assess officer performance in early 1900s. The practice was then translated to the industry world when employers realized its potential to improve organizational effectiveness. A modern evaluation process, performance appraisal, found its way to the modern management system.

The psychologist's role is to help organizations to improve the effectiveness of performance appraisal. They assisted employers to develop a variety of appraising procedures and tools, such as global ratings and global essays, judgmental rank order, graphic- and trait ratings, forced-choice, and critical incidence (see Landy & Farr, 1980; Wiese & Buckley, 1998). More rating formats, such as behaviorally anchored scales, behavioral expectation scales and behavioral observation scales, were developed during 1960s and 1970s, and "this period saw rating format research reach its zenith in popularity along with research on rater training and rater individual differences" (Austin & Villanova, 1992, p. 853). Meanwhile, the content of appraisal evolved from one global measure to several broad traits, from several broad traits to a number of specific traits, from traits to behaviors and short-term goals (see McGregor, 1957). Many studies were conducted to understand the relative strengths and weaknesses of these formats. Drawing on research on psychological ratings (e.g., Thorndike, 1920; Bingman, 1939; Knight, 1923; Moore, 1937), performance appraisal research during this period focused on rating bias due to raters and rating

formats. The latter has remained as one of the main topics of interest among researchers until today.

As more and more organizations accepted performance appraisal practices, some pioneers, either practitioners or researchers, began experimenting with using raters other than supervisors (e.g., Hollander, 1954, 1956, 1957; Maloney & Hinrichs, 1959; Rupe, 1951; Wherry & Fryer, 1949; Williams & Leavitt, 1947). Researchers in 1950s and 1960s were intrigued by these new forms of rating. For example, after peer-rating was introduced (Hollander, 1954), research was conducted to understand its psychometric qualities (e.g., Gordon & Medland, 1965; Fiske & Cox, 1960) and applications (Weitz, 1958). Later on, more and more studies focused on contrasting between ratings from different sources. For example, Lawler (1967) examined self-, supervisor-, and peer-ratings of performance with convergent and discriminant analyses and found that supervisor- and peer-ratings had greater validity than self-ratings. Thornton (1968) compared self-ratings with supervisor-ratings in predicting promotability, and identified some meaningful discrepancies between two perspectives. Landy and Farr (1980) described additional studies.

While research on different rating sources were receiving more and more attention in the industry world in late 1970s and 1980s, research took on a different direction and focused more on understanding the cognitive aspects of performance appraisal. Landy and Farr (1980) updated the process model of performance rating and included cognitive processes as a key component that

deserved more research. Feldman (1981) proposed a framework of cognitive processes, which was later refined by Ilgen and Feldman (1983). The framework includes attention, categorization, recall and information integration processes. Feldman's social cognitive approach suggested a research shift from rater and rating format to rating process. Many similar models were proposed (e.g., DeNisi, 1996; DeNisi, Cafferty & Meglino, 1984; Ilgen & Feldman, 1983; Spool, 1978). From a different field, Lord and his associates also proposed their leadership perception models (Hanges, Lord & Dickson, 2000; Lord, 1985; Lord, Foti & DeVader, 1984).

Cognitive approaches to performance appraisal, however, have some major problems. First, it is usually conducted in the laboratory and uses students as samples, thus its implications for industry are vague (Banks & Murphy, 1985; Bretz, Milkovich & Read, 1992; Napier & Latham, 1986). Second, while it focuses mainly on processes, social contexts are not adequately considered (Bernardin & Villanova, 1986). Feldman (1981) suggested that research first be conducted in the laboratory and then in the field, but research in the field often lagged behind. Banks and Murphy (1985) cautioned that the research-practice gap was widening and the cognitive processing approach failed to address the *willingness* component of rating, which clearly ties to rating contexts. In contrast, MSF incrementally gained popularity in the business world. Based on the research on multiple rating sources and the ideas of *assessment for development*, Robert Bailey from the Center for Creative Leadership developed the first MSF survey in

mid-1970s (see Campbell, 2001). Since then, MSF offerings grew steadily. Van Velsor and Leslie (1991) listed 16 MSF instruments and the number grew to 24 seven years later (Leslie & Fleenor, 1998). There are even more instruments owned by companies who have developed them privately. At the beginning of 1990s, MSF became a popular management practice in the business world and research using MSF data also increased dramatically afterwards.

There are two other issues in the field of performance appraisal and MSF that are worth mentioning; one is practice-related and the other is research-related. The first issue is about which purpose (administration or development) performance appraisal or MSF should serve. Performance appraisal or MSF was initially designed for administrative purposes, such as compensation, promotion, and so on. As an example, Weitz (1958) published an article labeled “Selecting supervisors with peer-ratings.” Performance appraisal for individual development started gaining attention in 1960s (read the Perce Corp story, Campbell, 2001). The 1970s and 1980s saw an intense debate over which purpose should take the priority or whether they could coexist. For instance, Reinhardt (1985) noted that performance appraisal used for administrative purposes still dominated in mid-1980s, while Campbell and Lee (1988) suggested using self-ratings for developmental rather than evaluative purposes. This debate ended up separating MSF from performance appraisal. The title now becomes “Don’t tie 360 feedback to pay” (Coates, 1998).

The other issue is about the relationship between different fields of research. From 1950s to 1970s, Mainstream researchers moved away from studying accuracy and focused more on rating bias, after Cronbach published his critique of accuracy research in person perception (Kenny & Albright, 1987). A small pocket of researchers, mainly I/O psychologists, continued to pursue accuracy, while others moved away from accuracy research and focused on a new arena—cognitive processes (Funder, 1995). Being away from the mainstream makes performance appraisal research a relatively closed field. Performance appraisal researchers seldom look to other literatures, such as selection and personality. Particularly, few MSF researchers notice person perception and interpersonal perception research (e.g., Funder, 1995; Kenny, 1991; Kenny, 1994; Kenny & Albright, 1987), although the relevance is high.

MSF in a Flat World

Pioneer users of performance appraisal, either ancient Chinese governments or modern American Army, are characterized by strict hierarchies. Before 1950s, bureaucratic structure and belief in rationality were also characteristics of business organizations (Ouchi & Wilkins, 1985). A top-down approach of performance appraisal is a natural extension of a paradigm that emphasizes objectivity, structure, and control. However, informal aspects of organizational life have received more and more attention since1950s, with different management paradigms being proposed (e.g., human relations, socio-

technical systems). Work groups and participation have become routine practices of more and more organizations since 1970s.

The organization is getting flatter as traditional hierarchical structure goes out of fashion and the team structure gains currency (Lawler, Mohrman & Ledford, 1995). Performance appraisal becomes a challenging task for supervisors in the new organizational reality. First, supervisors now have less opportunity to observe each person's performance due to increased span of control. Second, roles and responsibilities have been broadened with increasing interdependency among team members. Measures of performance have gone beyond pure job-related performance to include contextual performance (Borman & Motowidlo, 1993). Third, jobs have become more complex as they are more knowledge-based (Drucker, 1993) and supervisors may not know exactly what each employee is doing. Finally, a team-based structure and a participation-oriented culture acknowledge a need to balance company and individual interests. The development of MSF is contingent on these changes taking place in the organization. For instance, McGregor (1957) and Thornton (1980) noted that increased use of self-ratings can be attributed to the general trend toward participative management. Lawler (1994) also ascribed the use of different rating sources to the movement towards employee empowerment and team-based organization.

Globalization and the technology revolution also are making the world flatter (Friedman, 2005). Improved communication technologies and globalized

business practices make international virtual teams a way of living. People who are stationed in different parts of the world can now work closely as a team, even though they may never meet each other face-to-face. Collecting subordinates' performance information becomes particularly difficult for supervisors, as they may not have enough opportunity to observe those subordinates working in different locations or nations. Relying solely on the supervisor to conduct performance appraisal is virtually impossible. Moreover, subordinates may come from different cultures and have different ways of doing things, and supervisors who are not familiar with those cultures will have difficulty in rating their performance. As the world is getting flatter, performance appraisal needs to address these unprecedented challenges from geography and culture.

Summary for This Section

MSF originates from the traditional performance appraisal but is different from it. MSF is built on decades of performance appraisal research, with generations of psychologists' relentless efforts. Performance appraisal made its way to the business world after its initial success in the military. Many years' research on rating formats, rater characteristics, and rating processes has built up a strong scientific foundation for performance appraisal. However, research on alternative rating resources (self, peers, and subordinates) took a critical leap in mid-1970s and gave birth to MSF, which finally gained its independence from the traditional performance appraisal as performance appraisal research shied away from business needs. Social and technical forces also play an important

role in shaping MSF. Working for a flatter organization in a flatter world poses many unprecedented challenges to traditional performance appraisal. MSF has several advantages over traditional one-way performance appraisal. First, it helps to gather as much as information possible, despite geographical barriers. Second, it makes the workplace more participative. Third, it fosters a learning culture in the team and promotes employee development. Fourth, it increases the alignment among team members and with the organization. These strengths make MSF a powerful management tool in the new era.

States and Issues of MSF

The previous segment introduced some historical roots and current forces that contributed to MSF's development. MSF features multiple perspectives. As a result, its accuracy depends on respective reliabilities and validities of ratings provided by different sources. Among different rating sources, peer-ratings and supervisor-ratings are the most reliable, with the mean single rater reliability for supervisors being .52 and that for peers being .42 (Viswesvaran, Ones & Schmidt, 1996). Another meta-analysis done by Conway and Huffcutt (1997) reported slightly lower interrater reliability coefficients for supervisor and peer-ratings (.50 for supervisors and .37 for peers). The mean reliability to subordinate-ratings is lower. According to Conway and Huffcutt, the mean reliability for subordinates is .30. Although reliabilities for peer and subordinates appear quite low, increasing the number of raters can substantially improve the reliabilities. Conway and Huffcutt inferred that two supervisors or three peers

could boost the reliabilities to .70. For subordinates, as many as six would be enough to ensure a reliability of the same magnitude. These reliability coefficients are generally satisfactory. However, self-rating has been shown susceptible to self-inflation and have lower reliability and lower validity (e.g., Campbell & Lee, 1988; Dunning, Heath & Suls, 2004; Fox & Dinur, 1988; Thornton, 1980), possibly due to egocentric bias, which may result from ratee's defensiveness, self-esteem, and attribution propensity (Harris & Schaubroeck, 1988). Despite its lower reliability, self-rating offers a unique perspective to performance appraisal (Bassett & Meyer, 1968; Thornton, 1968). Moreover, self-rating plays a critical role in understanding self-other agreement.

While the reliability of MSF is generally adequate, the evidence for its validity is more controversial. The fundamental underlying assumption of MSF states that ratings from different sources (e.g., superiors, peers, subordinates) provide unique and meaningful information about a target manager's performance (Murphy & Cleveland, 1995), and thus a source effect will present and MSF will demonstrate higher validity than traditional one-way performance appraisal. There is some support for this notion. Barrett (1966) argued that organizational position of the rater has significant impact on the rating as people in different positions tend to have different opportunity to observe the ratee. Subordinates may be the best source of information on delegation, motivating and developing others while peers may be in the better position to rate competencies like decision making and collaboration (Pollack & Pollack, 1996).

Salam, Cox, and Sims (1997) also found that superiors and subordinates tended to perceive the performance of target managers differently, depending on their positions and organizational contexts. Pulakos, Schmitt and Chan (1996) observed discrepancies between supervisor-ratings and peer-ratings, and they also speculated that supervisors and peers might consider different performance factors when they provided ratings to the ratee.

However, Borman (1997) indicated that such differences are generally weak. Yammarino (2003) identified three key themes that contradict the assumption. First, rating source effects are no stronger than individual rater effects. Second, ratings between sources show about as much convergence as do ratings within sources, and their relationships to the latent construct are very similar. Third, multisource ratings may inform us more about the rater providing the data and his or her views rather than about the focal manager who is being rated and his or her actual performance. Greguras, Robie, and Born (2001), based on the social relations model, also demonstrated that how individuals see themselves is congruent with how they see others. Hogan and Shelton (1998) pointed out that we tend to think about people from our own perspective and rarely do so from their perspective, thus ignoring their goals, aspirations, and fears. If Yammarino, Greguras et al., and Hogan and Shelton are right, MSF should be repositioned as a tool that doesn't necessarily reflect the target manager's actual performance.

From the psychometric perspective, which presumes existence of a true performance, ratings from different sources are expected to have moderate correlations, based on the assumption that raters from different sources should contribute additional unique perspectives to the ratee's performance. However, most correlations are below .30s, except that the correlation between supervisor and peer reaches .34 (Conway & Huffcutt, 1997). Results from an earlier meta-analysis (Harris & Schaubroeck, 1988) seem more optimistic. The correlation between supervisor and peer is as large as .60, while the correlations between self and supervisor, self and peer are in mid .30s. The discrepancies, however, could result from the way Harris and Schaubroeck computed the correlations. James (1982) has shown that using composite scores could inflate correlation between two variables. The results of these two studies should come closer if the latter used single rater scores. Some moderators, such as job type, play a part in the cross-sources relationships. Conway and Huffcutt (1997) found that the supervisor-self and peer-self correlations in managerial jobs are substantially lower than those in non-managerial jobs, but the supervisor-peer correlations are not affected by job type. Similar patterns were found in the Harris and Schaubroeck (1988) study. Lower cross-source correlations in managerial jobs could result from lower reliabilities of ratings in these jobs, which in turn could be caused by job complexity.

Another psychometric approach is to break down the variance of ratings into components (usually, ratee, rater, ratee X rater, and measurement error) and

see whether the true performance, or the ratee effect, accounts for a large proportion of the variance. The theory of measurement (Wherry & Bartlett, 1982) provides a useful framework to conduct construct validity studies in MSF research as it specifies more subtle effects. The ratee effects are further separated into sub-components that correspond to the ratee's general performance and area/dimension specific performance. The rater effects are also broken down into sub-components that are source-specific and idiosyncratic to the rater. MSF assumes that using raters from different sources can maximize the ratee effects and minimize measurement errors. Lance (1994) found support for both true performance and idiosyncratic rater effects, with the latter accounting for the majority of the variance. Recently, Scullen, Mount and Goff (2000) used a two-step structural modeling process, Correlated Uniqueness-Confirmatory Factor Analysis (CU-CFA), to investigate the proportions of variance of various components. Results based on two large data sets (n = 2,350 and n = 2,142) and two forms of questionnaires showed that idiosyncratic rater effects accounted for over half of the rating variance (62% and 53% for two data sets, respectively), while random measurement error effects were 11% for the first data set and 18% for the second data. The source effects accounted for 9% of the variance in the first data set and 7% in the second. The results of this study are consistent with some researchers' (e.g., Greguras & Robie, 1998; Greguras, Robie & Born, 2001; Lance, 1994; Yammarino, 2003) speculations that raters see others as they see themselves. It is also consistent with findings

from the Mount et al. (1998) study, which demonstrated that the 10-factor model, representing 7 rater factors and 3 trait factors, had the best fit. The good news is, though, the measurement error is fairly small, while other effects account for a moderate amount of the variance. Scullen et al.'s study also lends some support to several assertions in the MSF field: (a) Multisource ratings reflect the ratee's performance; and (b) raters from different sources provide unique perspectives. However, multisource ratings don't seem to account for more performance-related variance than supervisor-ratings only.

Therefore, evidence of validity for MSF is quite mixed. While ratings by supervisors, peers, and subordinates are reliable, correlations among them are generally low. The fundamental assumption of MSF which states that each rating source will provide a unique perspective doesn't completely hold based on available research results (Greguras & Robie, 1998; Greguras, Robie & Born, 2001; Lance, 1994; Mount et al, 1998; Yammarino, 2003). Idiosyncratic rater effects are predominant while true performance accounts for only a small proportion of the variance. Scullen et al. (2000) did observe some source effects, but they also account for only a small proportion of the variance. Bearing all these negative messages in mind, should we continue using MSF? Three different discussions may help to answer this question. First, epistemologically, it can be argued that MSF is intended to measure different realities, so it is valuable if multiple realities can be accurately reflected. A description of two epistemological arguments will be offered. Second, a comparison between MSF

and assessment centers is conducted to identify improvement opportunities for MSF. Third, person perception research is briefly reviewed and its implications to MSF are discussed.

Epistemology of MSF

After reviewing performance appraisal literature on accuracy, Heneman, Wexley and Moore (1987) concluded that most research was constrained by theoretical and methodological limitations and laboratory settings, and thus offered few useful suggestions to the field. They went on saying, “Instead, accuracy may be the amount of interrater agreement between parties in the performance appraisal process as to the dimensions and level of performance. If this alternative conceptualization were adopted, this will open up a number of new avenues for future research” (p. 443). Heneman et al. posed an important question about whether a true performance score exists and what standards we should choose for accuracy study. The distinction between the realistic paradigm and the phenomenal paradigm (Kruglanski, 1989) is relevant here. The former term presumes a truth beyond perception while the latter term believes the truth is co-created by people in interaction, thus representing a constructivist perspective.

Such a distinction has important implications for MSF research. The psychometric approach represents the realistic paradigm. It assumes there is a true performance score and deviations from it are errors. In Wherry and Bartlett’s (1982) language, the ideal situation is that idiosyncratic rater effects and random

measurement errors are at a minimum. As shown in precedent research based on the theory of measurement (Wherry & Bartlett, 1982), while random measurement errors are low, idiosyncratic rater effects are prominently high. Nevertheless, the psychometric perspective dominates the field of performance appraisal (Ilgen & Feldman, 1983). Consistent with the phenomenal paradigm, Tornow (1993a) speculated that different sources could perceive different realities and thus have differential perceptions. Moses, Hollenbeck and Sorcher (1993) also suggested a need to differentiate between other people's observations and other people's expectations. The former is aligned with the psychometric perspective, while the latter puts more emphasis on social judgments, inferences, or expectations of how a person will behave in a given situation. Ashford (1989) also considered the organizational reality as socially constructed and thus claimed that both objective and subjective criteria are warranted in MSF research.

A similar debate takes place in the field of organizational climate research in the form of unit of theory. Organizational climate is a higher-level construct that reflects organizational attributes. However, organizational climate often bases its measurement on individual perceptions of organizational attributes. Individual perceptions are aggregated to represent organizational climate. From a realistic point of view, aggregation makes no sense as individual perceptions may depart from each other. To resolve this unit-of-analysis problem, James and Jones (1974) made a distinction between psychological climate and organizational

climate, treating the former as an individual-level construct and the latter as an organization-level construct. Scores from different individuals can either be aggregated into composites to represent organizational climate or remain as indicators of psychological climate (Jones and James, 1979). James (1982) emphasized the importance to make sure interrater agreement is high enough for aggregation. James used two climate variables: fairness and support to illustrate that to aggregate adequate interrater agreement must be present as there is no *a priori* assumption of a true score for fairness and support at the organizational level. When the item "treat others fairly" is rated by different individuals, the ratings could be very different and no true score can be determined. In this case, there is no support for the existence of a true reality.

As an example, Shamir, Zakay, Breinin and Popper (1998) demonstrated that supervisors and subordinates did have different expectations of a target leader's charismatic behaviors. In their study, supervisors in the military were rated higher in the leader's effectiveness when he/she displayed exemplary behaviors or emphasized ideology, while subordinates demonstrated stronger positive culture and discipline when the leader displayed behaviors that emphasized collective identity. Supervisors tended to pay more attention to the leader's alignment behaviors while subordinates seemed to enjoy unique unit identity more. In addition, Shamir et al. (1998) found that unit culture and discipline accounted for additional variance in supervisor-ratings beyond the leader's individual behaviors. Supervisor-ratings thus are based on both

leadership behaviors and unit-level products resulted from those behaviors. Shamir et al.'s study clearly demonstrates the multiple realities exist and aggregation provides additional information about leader perception. This example also shows MSF can not only be a tool for measuring, it can also be a tool for managing (Wiese & Buckley, 1998).

Heneman et al.'s (1987) sentiment that accuracy might not be the only criterion for performance appraisal research is worth noting. Instead of treating discrepancies in ratings as errors, other researchers (e.g., Heneman et al, 1987; Moses et al., 1993; Tornow, 1993a) suspect that multiple realities may exist. Performance appraisal and MSF should go beyond a narrow psychometric view (Ilgen & Feldman, 1983). A broader view of performance appraisal and MSF is warranted. In the practice side, MSF has already been used for organizational diagnosis (e.g., Church, Waclawski & Burke, 2001). A broader view also means to look beyond the MSF field and incorporate best practices and research conclusions from other research areas.

Comparing MSF with Assessment Centers

Assigning meaning to other's or one's own behavior is part of human nature, but it is psychologists who structure the process by designing scales and assigning numbers to psychological characteristics. The act of making judgments and assigning numbers according to a structured scale is called rating. There are two basic forms of rating: self-rating and other-rating. Traditional performance appraisal uses other rating predominantly but self-rating was introduced due to

its unique perspective (Bassett & Meyer, 1968; Thornton, 1968). In contrast, self-rating dominates personality testing but Epstein (1979, 1980) has proposed using other-rating to measure personality.

This section will focus on other-rating and make a comparison between MSF (also performance appraisal) and the assessment center. An assessment center features multiple assessors observing behaviors of a participant in multiple exercises on multiple dimensions (Thornton, 1992). There are various types of other rating (e.g., the structured interview, the assessment center, and the survey), among which the assessment center is the most structured and sophisticated. The assessment center is also a valid assessment tool (Hunter, Schmidt, & Jackson, 1982), especially when it is used for identifying potential and when the assessors are psychologists (Gaugler, Rosenthal, Thornton, Bentson, 1987). A comparison can help us to understand what the assessment center has done to improve accuracy and what MSF can learn from it (see Table 1).

Specifically, an assessment center uses multiple well-trained assessors to observe a participant. These assessors can be industrial psychologists or managers in the organization. To be qualified, managers are usually two levels above the participants, and they must receive training and be certified before conducting the center (Thornton, 1992). Such practices can not only make sure managers meet the requirements of assessors but also minimize possible role conflicts between assessors and participants. As a highly complicated assessment tool, extensive training is a prerequisite for assessors. Rater training

has been found to be effective in reducing rating errors in performance appraisal, especially when the training is extensive (Landy & Farr, 1980; Spool, 1978). Woehr and Huffcutt (1994) reviewed 29 studies in the area of rater training and found rater training improves rating accuracy. Particularly, Frame of Reference training produces an effect size of .83. MSF usually doesn't have a large rater pool from which most competent raters can be chosen, but rater training should be conducted consistently (Feldman, 1981).

Assessment centers measure complex behaviors that take place in a well-structured but highly complicated situation. To make observation and rating possible, assessors are instructed to focus on behaviors. Each exercise covers a certain amount of competencies, with detailed descriptions of behavioral standards. These standards represent different levels of behavioral effectiveness, with some typical behavioral examples. The assessor's job is to match behaviors they observed during a simulation with pre-established standards. With extensive training and experience, assessors will develop a well-structured cognitive map of behaviors and competencies. According to Ostroff and Ilgen (1985) and Ilgen, Barnes-Farrell, and McKellin (1993), consistent cognitive categories in the rater's mind, which are formed through training and experience, can help to improve rating accuracy. When the rater's schemas are consistent with the dimensions of performance, accuracy increases (DeNisi & Summers, 1986; Lauterbach & Barnes-Farrell, 1989). Mount and Thompson (1987) also found that subordinates' preconception of ideal supervisor behaviors influences their perceptions of real

Table 1

A Comparison between Assessment Centers and PA/MSF

Factor	Assessment Center	PA/MSF
Assessor/Rater	Multiple well trained and experienced assessors (psychologists, HR personnel, or managers two levels up)	Untrained or partially trained managers; for MSF, multiple raters are available
Content	Behaviors are grouped into competencies, with clear behavioral descriptions	Behaviors are less well organized, usually without clear behavioral descriptions
Process	Structured materials, processes, note-taking, behavior classification and integration	Give ratings based on recalled behaviors or general impressions; no classification or integration
Timeframe	1-3 days; rely on short-term memory	Usually one year; rely on long-term memory
Context	Context-free, non-workplace, for managers two levels higher, for psychologist usually a one-time interaction	Embedded in current organization, workplace activities, dyads, past and future interactions

supervisor behaviors. Although MSF is behavior-based, the behaviors are less well organized and specified than those in the assessment center, thus leaving some room for the rater's discretion.

To cope with information overload, a frequently encountered situation in assessment centers, assessors use notes to record behaviors. After each exercise, all behaviors recorded will be classified into different dimensions. Assessors then assign ratings to dimensions based on observed behaviors either before or after discussing their observations with partners. Finally, assessors gather together to discuss their ratings on different exercises and assign a final rating to each dimension. This well-sequenced process can ensure all relevant behaviors are recorded, correctly classified, and appropriately rated and integrated. In contrast, note-taking in performance appraisal and MSF is not a common practice, although there is an obvious reason for note-taking—recalling behaviors that took place long time ago are difficult to remember. MSF also uses multiple raters, but these raters are not supposed to pool their observations together or discuss their ratings publicly. Therefore, a consensus meeting or integration section is usually not available in either performance appraisal or MSF. It should be noted that the classification process is not usually present in performance appraisal or MSF.

Performance appraisal and MSF have a much longer timeframe than assessment centers. Assessment centers usually take 1-3 days. Assessors rely more on short-term memory and notes to organize behaviors and give ratings.

With no exception, assessors develop some general impressions about a participant too, but these impressions can be easily affirmed or refuted by discussion or the behavioral evidence at hand. Performance appraisal and MSF have a much longer timeframe, so raters have to rate behaviors that have happened up to one year before. DeNisi, Cafferty and Meglino (1984) noted that the tendency to recall impressions rather than behaviors will increase over time, as raters are more likely to store impressions than behaviors. Particularly, in situations of time constraints, cognitive overloading, and insufficient information, raters will resort to general impressions formed over the course of numerous interactions with the ratee. Such general impressions are difficult to overcome (Feldman, 1981). However, research has shown that the relationship between rating accuracy and time delay is quite complicated. Heneman and Wexley (1983) observed that delayed ratings were less accurate than immediate ratings, but several studies (e.g., Barnes-Farrell & Couture, 1984, cited in Ilgen et al., 1993; Sanchez & De La Torre, 1996; Sulsky & Day, 1994; Woehr & Huffcutt, 1994) have indicated that a delay doesn't necessarily mean lower rating accuracy, although a delay makes recalling behavior less accurate. These studies echo DeNisi et al.'s (1984) speculation that in the delay situation, raters rely more on their impressions than behavioral memories. Accurate behavior recall may be more important to assessment centers as immediate ratings are required (Barnes-Farrell & Couture, 1984).

One important feature of assessment centers is their high fidelity. Through simulations participants can demonstrate their past job-related behaviors. While assessment centers provide rich contextual information about simulations, they endeavor to minimize social or political influences from current employers. Ideal assessors are external psychologists, HR personnel, or managers two levels above participants. When mangers one level up must be included, they are not allowed to assess their direct reports. These arrangements can greatly minimize the impact of participants' past performance and past interactive experience with the assessors. In some cases, assessment centers are conducted off-site to avoid possible disturbances. Assessment centers thus can provide a safe and fair environment for both assessors and participants. However, performance appraisal and MSF embed themselves in rich social contexts. The rater and the ratee have been with each other for some time and they will remain with each other in the future. Contextual factors, such as quality of leader-member exchange (Heneman, Greenberger, & Annonyuo, 1989), expected future interactions (Axelrod, 1984; Ilgen & Favero, 1985), task characteristics (Klimoski & Donahue, 2001), and national culture (Atwater et al., 2005), may alter the rater's ratings.

Table 1 shows that MSF and the assessment center are different in many aspects, but this study is more focused on contextual factors. Interestingly, neither MSF nor assessment center research pays much attention to contextual factors, but for different reasons. Assessment center research has

disproportionately concentrated on assessor, content, and process, and generally ignores contextual factors, with a belief that they have been controlled. There are few studies that take cross-cultural issues into account (Lievens, n.d.). One exception, Lievens, Harris, Van Keer, and Bisqueret (2003) studied a sample of European executives who were selected to work in Japan, and they found that these executives' performance in a group discussion exercise strongly predicted future performance ratings by Japanese supervisors, while a presentation exercise failed to make similar predictions. No theory was developed to explain such differences. In the same vein, Landy and Farr (1980) and Feldman (1981) strongly urged more research on contextual factors in performance ratings. Klimoski and Donahue (2001) reviewed many studies that focused on contextual factors' effects on person perception, and developed a process model that includes these factors. However, they didn't specifically discuss national culture. Levy and Williams (2004) developed a framework that includes numerous contextual factors, but not much work related to national culture is reported. To understand how contextual factors (e.g., national culture) affect performance ratings, person perception research provides important insights.

Person Perception Research

Ilgen and Favero (1985) suggested that the constructs and theoretical frameworks from person perception research should be valuable to performance appraisal research. An example is the leadership perception model developed by

Hanges, Lord and Dickson (2000). In Hanges et al.'s model, self-concept plays a pivotal role. Self-concept can be referred to a theory of world (Palermo, 1983), mental programs (Hofstede, 1998), or motivational mechanisms (Mitchell, 1983). According to Hogan and Shelton (1998), self-concept defines a person's goals, aspirations, hopes, dreams, and ideal self-image and determines one's strategies to get along, get ahead, and find meaning.

Building on their leadership categorization theory (Lord, Foti and DeVader, 1984), Hanges, Lord and Dickson (2000) developed a new connectionist model of leadership perception. The leadership categorization theory states that there are multiple leadership prototypes in peoples' minds and leadership perceptions are created through matching an observed leadership behavior to these prototypes and then retrieving the most relevant one. The theory provides a serial and static account of leadership perception. To reflect the paralleling and dynamic nature of the process, Hanges et al. (2000) applied the connectionist model to their new leadership perception model. Leadership prototypes now no longer have symbolic existence but exist in the flux of widespread connections between nodes. Some connections will be more heavily weighted due to experience and repetition. The new model is open to various individual and contextual influences. Cultural, structural and individual factors exert their influences through imposing *constraints* on self-concept. Lord et al. (2001) stated, "Thus, we propose that coherence in leadership perceptions comes from satisfying multiple constraints on prototypes that vary, depending upon factors

such as national cultures, organizational context, specific tasks, leader qualities, perceiver expectations and implicit theories, and immediate history" (p. 314).

Lord, Brown, Harvey and Hall (2001) went further to consider multiple level issues with their new model. They suggested that the most appropriate level of analysis depends on the strongest constraint on the process. When national culture becomes salient and acts as a key constraint, the output of the connectionist model will be culturally bound and a cultural level of analysis would be appropriate. When the dyad of leader and follower becomes salient and is conceived as the key constraint, a dyadic level of analysis would be appropriate. This multi-level conceptualization of leadership perception is aligned with the multi-level view of self-identity (Lord & Smith, 1999; Markus & Kitayama, 1991; Markus & Wurf, 1987; Shweder & Sullivan, 1993; Triandis, 1995). The activation of the individual, interpersonal and collective self-identity could produce results that opt for individual, dyadic, and group level analysis, respectively. Therefore, the most appropriate level of analysis depends on the interplay between external and internal constraints.

Consistency of perceptions across time and/or persons can be achieved if strong higher-level constraints are present (Lord et al., 2001). Consistency makes aggregation of perceptions across persons possible and a higher level analysis legitimate. Therefore, salient external constraints often times justify the act of aggregation. Interestingly, leadership behavior itself can be a salient constraint that justifies aggregation, because it could have a strong impact on

others. Leaders can directly influence subordinates' self identities, thus resulting increased consistency of perceptions among subordinates (Lord & Smith, 1999). Leaders can also influence subordinates indirectly through organizational or group climate that they foster (Lord et al, 2001). It should be noted that not all leaders are influential enough to change other's self-identity and for some organizations and groups a consistent climate may not exist. An individual or dyadic level of analysis would be more appropriate, provided that no other higher-level constraints are strong.

In sum, the connectionist model of leadership perception (Hanges et al., 2000) provides a useful framework to explain how leadership behavior is perceived. The network can be automatic. People may not be conscious of culture's role in the perceiving process. The network is dynamic. It shows how goals, values, norms, and affects which originate from different sources (e.g., culture, task, leader, follower) impose constraints on the process, and how these constraints influence consistency in leadership perceptions and thus define the appropriate level of analysis. Thus, the connectionist model can help to link extraneous contextual factors, such as national culture, to self-concept (Lord & Brown, 2004). Due to its dynamic properties and its capability to account for multi-level phenomena, the connectionist model will be adopted to create a new conceptual framework of self-other agreement.

Weighted-Average Model

While Hanges et al.'s (2000) connectionist model provides a framework that can be used to understand how external and higher-level constructs, like national culture, influence person perception through self-concept. Kenny (1991) proposed a mathematical model that simulates how factors of acquaintance, overlap, consistency, similar meaning systems, unique impression, and communication affect perceivers' consensus. Kenny (1994) expanded the model by including three factors that are related to stereotypes, but this study will not include them for the sake of simplicity. Kenny's models provide rich information about how these factors function together to influence interpersonal perceptions. Six concepts that appeared in Kenny's (1991) model will be introduced below.

Based on Kenny's (1991) weighted-average model, acquaintance predicts consensus in the first few interactions, but the relationship will level off quickly. This helps to explain King, Hunter and Schmidt's (1980) notion that the reliability of summated ratings for a single supervisor has a ceiling value of .60. Based on ratings of 9,975 ratees from 79 organizations, Rothstein (1990) observed a curvilinear relationship between rater's length of exposure to the ratee and interrater reliability, which reached to asymptote level at around .60, lending strong support to Kenny's projection. Rothstein also found that lower interrater reliability is related to smaller variance in ratings, possibly due to leniency, severity, or central tendency biases. When raters have not known the ratee long

enough, they may be less confident in rating his/her behaviors, and thus use more heuristics to fill the gaps.

Table 2

Definitions of Factors in Kenny's (1991) Weighted-Average Model

Factor	Definition
Acquaintance	The amount of information to which the perceiver is exposed
Overlap	The extent to which behaviors are observed by both perceivers
Consistency	The extent to which the target exhibits consistent behaviors across situations
Similar Meaning Systems	The extent to which an act is interpreted similarly by two perceivers
Unique impression	The extent to which perceivers rate the target based on extraneous information, such as perceiver's mood state
Communication	The extent to which perceivers share their perceptions of the target with each other

However, such a moderating effect of acquaintance didn't show up in the Bernieri, Zuckerman, Koestner and Rosenthal (1994) study. Length of acquaintance was found to be unrelated to the degree of consensus between self- and other-ratings of personality traits. Anyhow, this is not unusual when a

dichotomous measure is used to predict a curvilinear relationship. In addition, Bernieri et al. did observe a moderating effect of cohabitation on personality ratings. The longer the cohabitation, the higher the consensus between self- and other-ratings. This finding substantiates Kenny's projection that overlap has strong impact on consensus. Long cohabitation gives the rater an opportunity to observe a wide array of behaviors, and thus results in higher consensus. Bernieri et al. noticed that the cohabitation effect was more prominent for girls than boys. This makes intuitive sense as girls in cohabitation usually share more social activities together (e.g., shopping) than boys do. Interestingly, the assessment center is structured in a way that multiple assessors observe the same set of behaviors. According to Kenny, acquaintance is not necessary for consensus when overlap is high. Kenny also predicts that when communication between raters is not allowed, the maximum limit of rater agreement is constrained by similarity between raters. When communication among raters is allowed, rater agreement can increase beyond the limit set by rater similarity. Thus, the practice of consensus meetings in assessment centers helps assessors to reach consensus.

Funder, Kolar and Blackman (1995) investigated the effects of acquaintance, overlap, similarity in meaning systems, and communication. Their study showed that raters do not need to be in the same context to agree with each other on personality ratings measured by Costa and McCrae's (1985) NEO Personality Inventory (NEO-PI), although being in the same context could make

ratings from different raters agree more with each other. Similar ratings do not necessarily mean that raters have to observe the same behaviors. Funder et al. also found evidence for implicit personality theory, i.e., raters give ratings partly based on their own self-images. Therefore, similarity among raters could increase agreement in ratings. Funder et al. believed the similarity effect is prominent in real life as similarity among raters usually is generally low. Nevertheless, some clusters based on geographical, ethnical or socioeconomic proximity cannot be ruled out.

Another important proposition of Kenny's (1991) model is that consensus and accuracy don't necessarily co-vary. When there exists a certain degree of behavioral consistency and the same behaviors are observed, the more behaviors observed the higher the accuracy, even when consensus has reached asymptote. If the proposition is correct, the validity of an assessment center would increase with more behaviors being sampled. Heneman and Wexley (1983) investigated the relationship between the number of behaviors observed and rating accuracy in the laboratory setting. They confirmed Kenny's proposition that the more behaviors were observed the more accurate the rating is. Kenny (1991) also proposed that communication helps to increase accuracy beyond asymptote, especially when the overlap parameter is low. This proposition, once confirmed, can also lend support to the integration section of assessment centers, where assessors come to share their observations of a participant in different simulations.

Kenny's (1991) weighted-average model has unique value with its strong emphasis on behavior. Kenny concluded that judges are indeed accurate in their perceptions of targets, and overlap and similarity of meaning systems have the strongest impact on consensus among all the factors. However, Kenny also considered other factors that might affect the usefulness of the model. For example, behavioral consistency might be inflated when ratings are influenced by prior acts of the ratee, or when behaviors are sampled from one single situation. In addition, acts could be differentially weighted so that first impression or most recent acts may have larger effects. Finally, when the timeframe is long, behavioral consistency will be lower, as behaviors tend to take place discretely. Due to these factors, the ideal weighted-average model may not apply in MSF. Nevertheless, the model offers a useful framework to discuss contextual factors' impact on consensus.

Summary for This Section

MSF has been demonstrated to be reliable but its validity is quite controversial. The assumption that MSF provides more valid data than supervisor-rating is not supported. MSF tackles a small proportion of true performance and shows weak level effects, while idiosyncratic rater effects are disproportionately high. Such results are disappointing to researchers who hold the realistic point of view, while those from the phenomenal camp who endorse socially constructed realities view such results more positively. A comparison with assessment centers shows that MSF generally does little to control

contextual influences and that more research on contextual factors is warranted. Research in other fields, such as leadership perception and person perception, provides some important insights into MSF research, particularly, to self-other agreement.

Self-Other Agreement

Past research has concentrated on how to improve performance appraisal accuracy by reducing rating bias, with the assumption that reducing biases increases accuracy (Ilgen & Feldman, 1983). Acknowledging the importance of accuracy research, Ilgen et al. (1993) conceived that other criteria, such as behavioral specificity and perceived fairness, are also important in practice. With a belief that the discrepancy view is more appealing, Ilgen et al. (1993) proposed a research shift "from rater errors to the discrepancy between ratings and some known standard of performance" (p. 353). However, they confessed that such a "known standard of performance" is difficult to define. From a self-assessment perspective, Ashford (1989) suggested that self-other comparison be a natural occurrence in a socially constructed organizational reality, as it provides important information concerning one's fit with the reality. Using other's opinions as an accuracy criterion thus is entirely appropriate. In this sense, self-rating is meaningful despite how inaccurate it is. Based on these notions, research on self-other agreement represents an important shift of focus in performance appraisal.

Self-other agreement represents the degree to which individuals see themselves as others see them. In order to maintain positive leadership perceptions effective leaders must stay attuned with their environment and adjust their behavior accordingly (Hanges, Lord & Dickson, 2000). Self-other discrepancy may indicate leaders are either out of touch with or insensitive to the

environment. It could be an indication of lack of self-monitoring and behavioral flexibility. Hogan and Shelton (1998) speculated that high self-monitors are more motivated to meet others' expectations and thus are more "rewarding" to others. A meta-analysis conducted by Day, Schleicher, Unckless and Hiller (2002) also suggests self-monitoring promotes job performance. Therefore, a positive relationship is expected between self-other agreement and performance.

From the psychometric perspective, MSF researchers (e.g., Fletcher, 1997; Wohlers & London, 1989) see self-other agreement as an index of self-awareness. Operationally, self-awareness can be defined as congruence between one's self-assessment of performance and performance viewed by other sources, such as supervisors, peers, subordinates, and customers (London & Smither, 1995). From the psychometric perspective, several conditions must be met to justify self-other agreement research. First, construct validity must be present. Second, the process of how the self and others reach agreement should be clarified. Finally, the relationship between self-other agreement and other variables (e.g., performance) can be predicted by the model.

Construct Validity of Self-Other agreement

Past reliability research has focused on either self- or other-ratings, instead of the self-other agreement measure itself. One exception is the Nilsen and Campbell (1993) study, which reported a test-retest correlation coefficient (30 days apart) of .71. Some support for construct validity is also present. Self-other agreement has been found to be related to measures of self-monitoring

(Church, 1997), locus of control and need for achievement (Mabe & West, 1982), and conscientiousness (Fletcher, 1997).

Mounting evidence has shown that self-other agreement can be used to predict performance (Alimo-Metcalfe, 1998; Atwater, Ostroff, Yammarino, & Fleenor, 1998; Atwater & Yammarino, 1992; Bass & Yammarino, 1991; Furnham & Stringfield, 1994; McCaulley & Lombardo, 1990; Van Velsor, Taylor & Leslie, 1993). The earliest support dates back to Thornton (1968). He studied supervisor-rating and self-ratings on 27 performance variables for the top executives of a large company and found that those executives who tended to overestimate themselves were least promotable. Bass and Yammarino (1991) also noted that their self-insight index of self-enhancement was related to promotion and supervisor-ratings of performance. Van Velsor, Taylor, and Leslie (1993) observed that over-estimators had the lowest performance, in comparison to accurate estimators and under-estimators. By using polynomial regression procedures suggested by Edwards (1993), Atwater, Ostroff, Yammarino and Fleenor (1998) found that effectiveness rated by supervisors was the highest when both self- and other-ratings were high, while effectiveness was the lowest when self-ratings were higher than other-ratings. However, Atwater et al. cautioned that such relationships might not hold when other criteria were used. Atkins and Wood (2002) used objective criteria - assessment center ratings - in their study. They found that while self-supervisor ratings predicted assessment center results, self-peer ratings failed to make similar predictions. Self-ratings

correlated negatively with assessment center results. Therefore, Atkins and Wood concluded that the relationship between self-other agreement and performance was inflated in the Atwater et al. (1998) study.

Another line of support is from personality research. Kwan et al.'s (2004) study developed a self-enhancement index based on Kenny's (1994) Social Relations Model (SRM). SRM can successfully separate perceiver effects, target effects, and their interaction effects, and thus can be used to study consensus among perceivers. The self-enhancement index was developed to represent the degree of consensus between self-ratings and other-ratings on a personality measure. Based on the data from 132 undergraduate students, Kwan et al. observed that their self-enhancement index predicted self-esteem positively. No relationship existed between the index and relationship harmony, a variable representing the extent to which individuals value harmonious relationship. More interestingly, the self-enhancement index had a moderate negative relationship with task performance that was objectively measured. They commented, "Self-enhancement bias may have positive consequences for intrapsychic adjustment, allowing self-enhancers to feel good about themselves (e.g., self-esteem, subjective wellbeing), but not for task performance and interpersonal adjustment (e.g., being liked), especially in longer term relationships when initially positive impressions may fade over time" (p.106). Overly positive self-evaluations may reflect poor social skills and maladjustment, which in turn have short-term and long-term negative consequences on mental health (Colvin, Block & Funder,

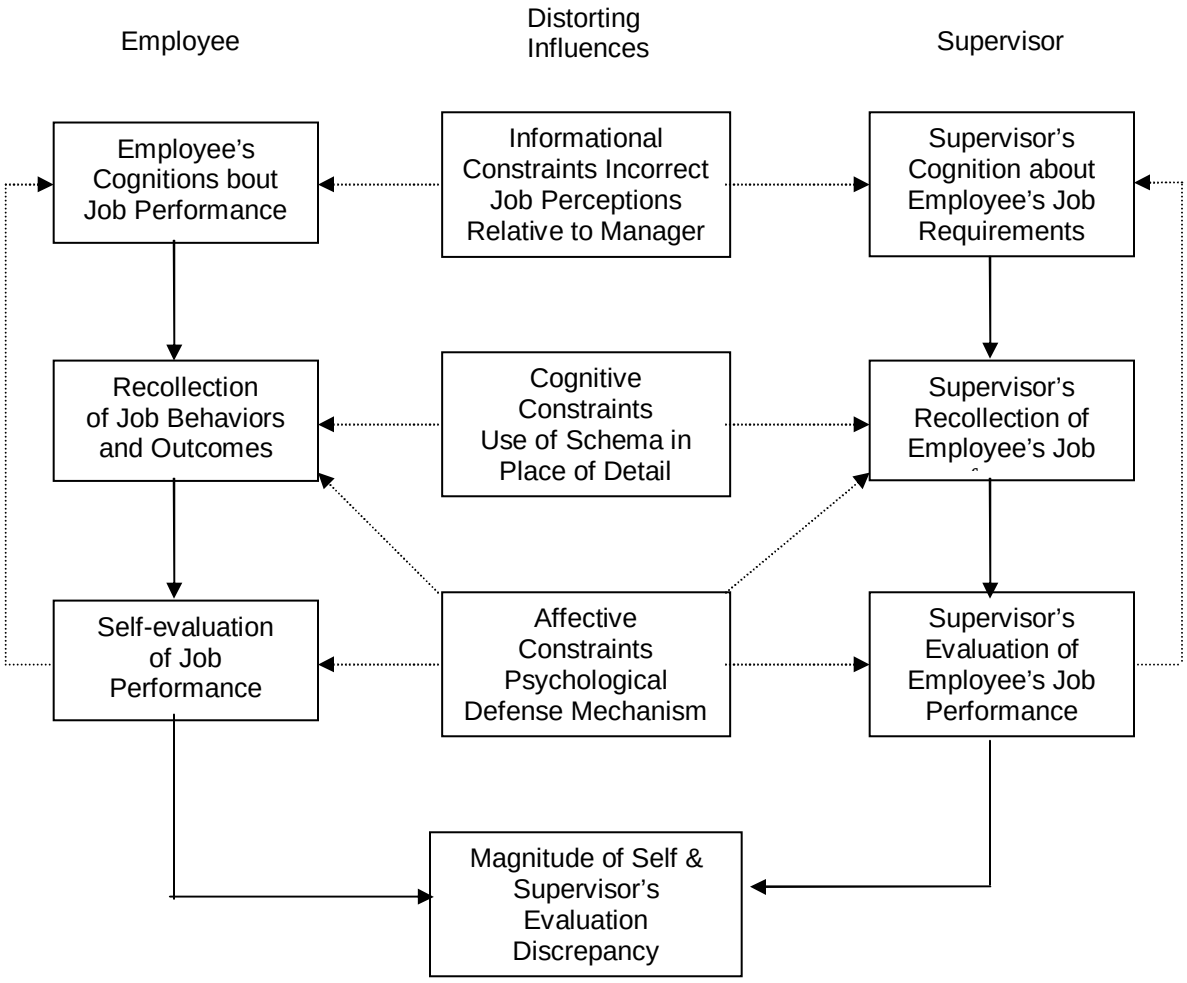
1995). Thus, the findings from the self-enhancement bias research are consistent with those from the self-other agreement research. However, it should be noted that the self-other agreement index is contaminated by the perceiver effects (Kwan et al., 2004). When strong perceiver effects are present, the relationship between self-enhancement and task performance will be attenuated.

Models of Self-Other Agreement

The previous review demonstrates that self-other agreement is a reliable and meaningful construct, but little theoretical work exists regarding how self-other agreement is achieved. As few exceptions, Campbell (1985) proposed a cognitive process model that portrays how information, cognitive and affective constraints lead to self-other discrepancies. Yammarino and Atwater (1993) proposed a well-known model of self-perception accuracy. Based on the degree of self-other discrepancies, they classified ratees into three categories: over-estimators, accurate estimators, and under-estimators. They suggested that these three categories of ratees would have different individual and organizational outcomes. While these models provide some practical guidelines to self-other agreement research, they fail to address many important issues. First, it is unclear how contextual factors affect self-other agreement. In addition, these models are static and linear, failing to account for the dynamism present in person perception. Finally, they have difficulty in addressing multi-level issues. To overcome these drawbacks, a new framework based on Hanges et al.'s (2000) connectionist model is proposed.

The Campbell (1985) model. Campbell (1985) proposed that discrepancies between employee self-ratings and supervisor-ratings come from three sources (see Figure 1). First, employees and supervisors may have different understandings of job requirements and goals. Due to information constraints, employees usually do not have a clear idea of what supervisors expect of him/her. Supervisors should clarify roles to employees in order to overcome the information constraints. Second, performance appraisal is a process that requires a lot of cognitive resources - to observe, recode, store, recall, and integrate behaviors. To reduce cognitive overload, behavioral information is usually organized in the form of schema and overall impression is retrieved when the supervisor evaluates the employee's performance. Discrepancies appear when the employee and the supervisor have different schemas and when they have different interpretations regarding the same behavior. Solutions to this problem can include rater training and adopting behaviorally anchored scales. Third, the impact of affect becomes salient when it gets to the evaluation stage. Subconscious motives may express themselves in forms of psychological defensive mechanisms or stereotypes, and the feeling of threat or guilt may play in. They color the cognitive processes of both the employee and the supervisor. One way to control this affect-based bias is to use the ratings for only one purpose. Information constraints, cognitive constraints, and affective constraints thus cause discrepancies between the employee's and the supervisor's ratings.

Figure 1. Campbell's (1985) self-other agreement model.



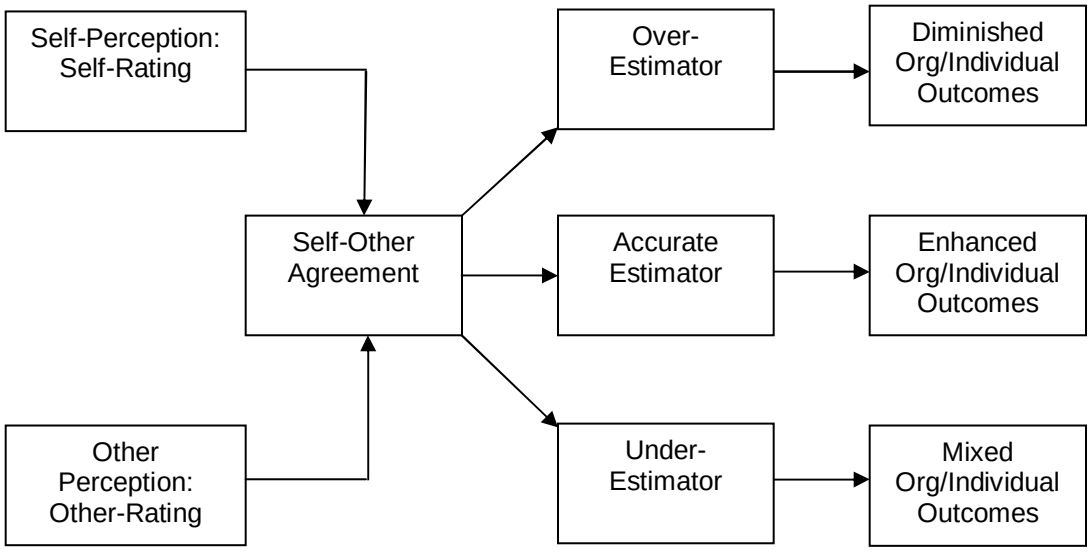
Note: Adapted from Campbell (1985)

In addition to the constraints, a feedback loop runs from "self-evaluation of job performance" to "employee's cognitions about job requirements" for the employee. When discrepancies exist, employees may choose to either put in more effort or lower their expectations. In parallel, a feedback loop goes from "supervisor's evaluation of the employee's performance" to "supervisor's cognitions about employee's job requirements" for the supervisor. When

discrepancies appear, the supervisor may either modify his/her cognitions about the job requirements or demand the focal employee to put in more effort. Campbell's model is straightforward and considers many possible causes of bias, such as attribution, psychological defensive mechanisms, and stereotypes. However, it pays more attention to intrapersonal factors, thus leaving contextual factors unattended.

The model of self-perception accuracy. Building on Ashford's (1989) framework of self-assessment, Yammarino and Atwater (1993) came up with a model of self-perception accuracy to describe self-other agreement and its outcomes (see Figure 2). They identified various factors that might influence self-ratings, other-ratings, and subsequently self-other agreement, but their model placed more emphasis on different outcomes resulting from self-other agreement. First, self-ratings and other-ratings are determined by self-perceptions and other-perceptions, respectively. Those factors that might influence self-perceptions and other-perceptions, including initial perceptions, individual characteristics, task characteristics, and the social environment, were identified and discussed. Second, self-ratings and other-ratings form an index of self-other agreement. Third, based on the degree of agreement between self-ratings and other-ratings, ratees are categorized into three groups: over-estimators, accurate estimators, and under-estimators. Fourth, different organizational and individual outcomes were suggested for these three groups of ratees.

Figure 2. Yammarino and Atwater's (1993) model of self-perception accuracy.



Note: Adopted from Yammarino & Atwater (1993)

Over-estimators are defined as individuals whose self-ratings are one-half of a standard deviation above other-ratings, accurate estimators are individuals whose self-other difference scores are within one-half of a standard deviation of other-ratings, and under-estimators are individuals whose self-ratings are one-half of a standard deviation below other-ratings. According to Yammarino and Atwater (1993), over-estimating will lead to diminished outcomes; accurate self perception may result in enhanced performance; under-estimators are characterized with mixed outcomes. A few studies (Atwater & Yammarino, 1992; Bass & Yammarino, 1991; Mabe & West, 1982; McCauley and Lombardo, 1990; Wholers & London, 1989) have found support for the model. Yammarino and Atwater also demonstrated that their categorization did not result from self-rating bias, as both self- and other-ratings play a part: Over-estimators have both

higher self-ratings and lower other-ratings, while under-estimators are characterized with both lower self-ratings and higher other-ratings (Atwater & Yammarino, 1992; Van Veslor et al., 1993).

The model of self-perception accuracy (Yammarino & Atwater, 1993) provides a general framework for MSF researchers to study self-other agreement. In comparison to Campbell's (1985) model, Yammarino and Atwater's model places more emphasis on possible outcomes of self-other agreement, which helps to stimulate a strong research interest in this field. However, Yammarino and Atwater's model has two obvious shortcomings. First, it does not delineate how self-other agreement is achieved. Second, the model considers contextual factors as antecedents of self-perceptions and other-perceptions, stead of the self-other agreement process. Atwater and Yammarino (1997) posted some updates on their original model. The most significant one was to separate accurate raters further into two groups: those with high performance (In-Agreement/Good) and those with poor performance (In-Agreement/Bad). Despite all these updates, the drawbacks that appeared in their original model remain.

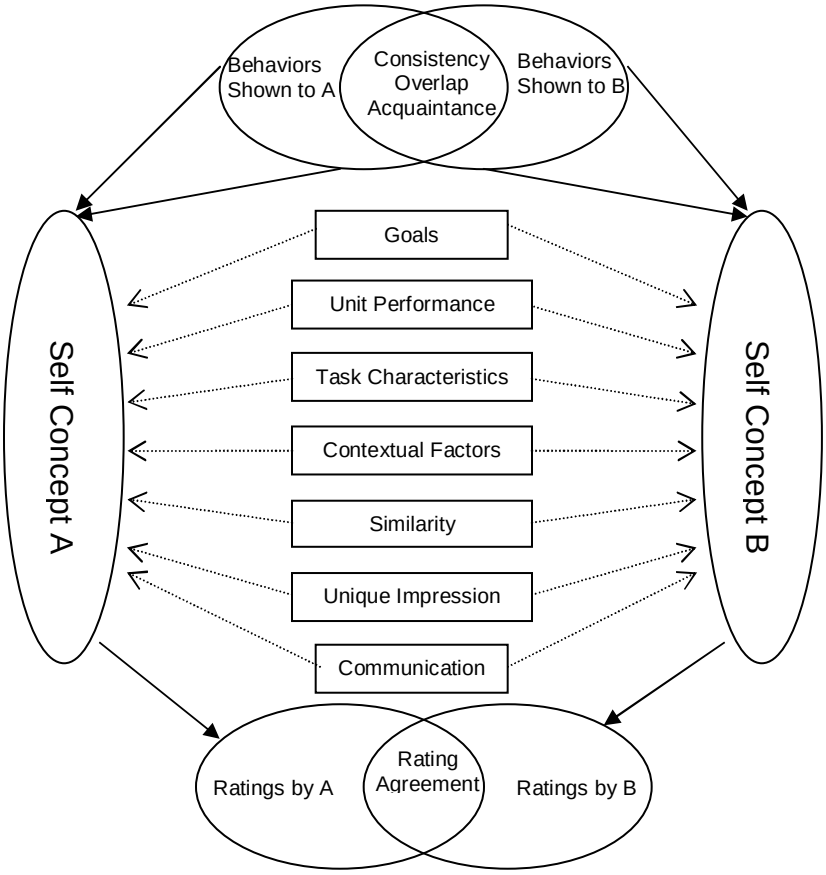
A new conceptual framework. Based on Hanges, Lord and Dickson's (2000) connectionist model and Kenny's (1991) weighted-average model, a new framework of self-other agreement is proposed (see Figure 3). The new framework will be able to overcome some shortcomings of the aforementioned models of self-other agreement (Atwater & Yammarino, 1997; Campbell, 1988; Yammarino & Atwater, 1993). First, the new framework catches the active,

dynamic, and nonlinear aspects of the self-other agreement process. Second, the new framework offers important insights into how various contextual factors affect self-other agreement. Third, the new framework provides a unique perspective to understand multi-level phenomena which have been generally neglected in MSF research. Finally, the new framework integrates many different theories, such cognitive theories, motivation theories, affect theories, and self-identity theories. The framework basically is a cognitive model that centers on self-concept. According to Schein (1990), a cognitive model can provide a useful vehicle to understand the process of meaning creation as well as how culture is learned. The new framework is generic in the sense that it can be used to account for self-other agreement by treating one of the raters as the ratee. The model has four elements: behavior, self-concept, rating, and constraint. A detailed description of these elements is presented below.

First, behavior. When a rater and a ratee come to know each other, their acquaintance depends on the frequency and the length of time they interact. Each dyad is unique. As different raters may interact with the same ratee in different circumstances, they will be exposed to different sets of behaviors of the ratee. As a result, a rater may observe only a subset of the ratee's behavior repertoire. In Figure 3, two intersecting circles are used to represent two behavioral sets. It should be noted that these two behavioral sets may not necessarily intersect, as two raters may observe totally different sets of behaviors. When the rater is the ratee himself/herself, he/she will have access to the whole

behavioral set in theory, while an external rater covers only a proportion of it. The way two circles intersect will be slightly different from that in the figure.

Figure 3. A new conceptual framework of self-other agreement.



Second, self-concept. Lord & Brown (2004) noted that self-concept plays a pivotal role in processing behavioral information and contextual information in the process of producing ratings. Self-concept is conceptualized as an active, dynamic, and nonlinear network-like structure in which many psychological processes are embedded. According to Lord and Brown, “In essence, the self is a collection of small, relatively independent processing units that are elicited in

different contexts and each of which has specific cognitive, emotional, motivational, and behavioral consequences" (p. 8). Researchers have shown that self-concept has different facets. For example, it can be individual, relational, and collective (Markus & Kitayama, 1991; Lord, Brown & Freiberg, 1999); it mirrors human beings' motives to get along, get ahead, and find meaning (Hogan & Shelton, 1998); it can be separated into self-views, possible selves, and current goals (Lord & Brown, 2004; Lord, Brown & Freiberg, 1999). The cognitive, motivational, affective processes that lead to rating biases and errors (see Klimoski & Donahue, 2001) can also find their habitants in self-concept.

Third, constraint. There are many contextual constraints but only a few will be discussed in the model. DeNisi, Cafferty and Meglino (1984) suggested that the purpose for appraisal has a strong influence on different aspects of the cognition process. Raters, like tacticians, can adopt different approaches to rating (Fiske, 1993). When the goal of rating is aligned between two raters, they may take the same strategy, and subsequently have similar ratings. The unit-level performance also affects performance rating. Keller (1992) suggested that team performance would have an impact on the perception of leadership behavior. In his study, team performance at Time 1 was found to influence team member's leadership ratings at Time 2. The degree of agreement in performance ratings is related to task complexity as well. It has been found that self-other agreement in managerial jobs is substantially lower than that for non-managerial jobs (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988). Lord et al. (2001)

suggested an indirect way for leaders to influence subordinates is through organizational or group climate that they foster. Strong organizational or group climate can make ratings from two raters more similar. While these contextual factors are important, this study will focus on national culture and the factors in Kenny's (1991) model: acquaintance, overlap, behavioral consistency, similar meaning systems, communication, and unique impression.

Fourth, rating. ratings are given by the raters based on the behaviors they observe. The extent of agreement depends on the behavioral sets raters are exposed to, differences in raters' self-concepts, as well as contextual influences. Agreement can be represented with two circles that intersect. The larger the intersection, the higher the agreement. When self-other agreement is the case, the ratee can be categorized into different groups using Yammarino and Atwater's (1993) model of self-perception accuracy and different outcomes will follow. As the main object of the current model is to further our understanding of how contextual factors influence self-other agreement, outcomes resulting from self-other agreement are not included in the model.

The framework shown in Figure 3 describes how two raters reach consensus on a leadership behavior as external observers. The process is quite straightforward, with the raters observing, processing, and giving ratings. When self-other agreement is the consideration, i.e., one of the raters is the ratee himself/herself, the process takes on additional complexity. The ratee can take three different approaches to assign himself/herself a rating. First, the ratee

observes or recalls his/her own behavior and its impact on others. People usually take this approach when there are ambiguities and they are not sure about the effects of their behaviors (Kenny, 1994). Second, the ratee rates himself/herself based on the feedback from others. People may also communicate directly with others and get to know how he/she is doing, when ambiguities arise. While this approach is helpful, people seldom ask for feedback from others (Kenny, 1994). Third, the ratee resorts to his/her self-concept as source of rating and doesn't have to recall specific behaviors. This happens when people have repeated the same behavior over and over again. Managers with a long tenure tend to over-estimate themselves (Ostroff et al., 2004), perhaps because they have formed stable self-perceptions over time and are not able to adapt their self-perceptions to the changing environment. It is also possible that ambiguities cannot be resolved by observing behavior or communicating with others. Correspondingly, there are three possible reasons for self-other discrepancies: the ratee is not able to correctly identify the impact of his/her own behaviors; the ratee reports self-perception directly from the self-concept without relaying it to behaviors; the ratee fails to get feedback from others. These reasons are present in both over-estimators and under-estimators (Yammarino & Atwater, 1993).

Overestimation may imply deficiency of social skills. Colvin, Block & Funder (1995) have suggested that overestimation reflects poor social skills and maladjustment. Over-estimators may also have an overly strong self-concept, as they tend to pay less attention to how others see them and are more interested in

maintaining a positive view of themselves by accepting positive feedback and discounting/rationalizing negative feedback (Yammarino & Atwater, 1997). As a result, they are more likely to derive ratings from their self-concepts rather than observe their own behaviors and ask for feedback from others. In comparison to over-estimators, under-estimators are more reserved. They have low aspirations and are more satisfied with their current states (Yammarino & Atwater, 1997). They don't necessarily have poor social skills (at least not as poor as over-estimators) but their self-concepts are quite negative. They are very sensitive to feedback from others but pay more attention to negative information, which is often consistent with their self-concepts. They tend not to ask for feedback proactively as they fear receiving negative feedback. Hence, both over-estimators and under-estimators may hold themselves back with confined self-concept and limited communication with others about their performance. Other factors may also play a part. For example, a ratee may overestimate to get external rewards or avoid punishments, or a ratee may just lack opportunities to observe the impact of his/her own behavior or seek feedback from others.

Summary for This Section

In sum, self-other agreement has demonstrated itself to be an interesting area for research. Self-other agreement is related to such constructs as self monitoring, self-awareness, and performance. Two models of self-other agreement (Campbell, 1985; Yammarino & Atwater, 1993) are introduced and a new conceptual framework is developed, based on the Hanges et al. (2000)

connectionist model of leadership perception and the Kenny (1991) weighted-average model. The new framework helps to overcome several limitations of the preceding models. In particular, it is very useful to understand how contextual factors, national culture in particular, influence self-other agreement. National culture as an important contextual factor has two roles to play in this model. First, it helps to shape self-concept and underlying psychological processes. Considering the pivotal role self-concept plays in the model and the evidence of cross-cultural differences found in self-identity research (e.g., Markus & Kitayama, 1991), it is reasonable to assume that national culture will make a strong impact on self-other agreement. Second, culture also affects collective behavioral patterns (Kitayama & Markus, 1999; Kitayama, 2002). Such collective behavioral patterns will in turn activate different aspects of self-concept and consequently affect self-other agreement. Before going any further, the construct of national culture will be introduced.

National Culture

Culture is an elusive construct that can be subject to different understandings. It can be something as profound as basic assumptions (e.g., Schein, 1990, 1992) deep in our mind, or something as vivid as routine rituals (e.g., Trice & Beyer, 1984). According to Wuthnow and Witten (1988), there are two traditions of culture studies, one focusing on explicit products of social interaction, another focusing on implicit features of social life. Culture is treated as something that is shared in a population. Special patterns and regularities in a

population, such as symbols, myths, and discourses, are seen as important vehicles through which culture can be studied. On the other hand, the implicit camp dives into people's minds and tries to find special patterns and regularities inside. Both camps hold a strong belief that culture is something that exists either *in the head* or *out there*, and thus can be labeled as the *entity view* of culture (Kitayama, 2002).

An alternative view, which is more constructivist but not as internally cohesive as other views, holds that culture is distributive in nature. A broad culture system includes external collective behavioral patterns, reifications, and internal psychological structures or processes. Culture is social and historical in the sense that people carry it from the past and transmit it to future generations collectively. Kitayama (2002) called this perspective the *system view* of culture. A more radical version of this system view is that culture is emergent, and is constructed through the "dialog" between the actor and external culture manifestations (Meek, 1988). The *system view* of culture thus indicates a need to consider both the historical roots of a culture and its current embodiments, and take both implicit and explicit features of social life into account.

Definition of Culture

Hofstede's (1980a) framework of national culture is used in this study, for two reasons: (a) it provides a succinct set of dimensions that can be used to compare different cultures; (b) it has resulted in a bulk of research that contributes to better understanding of national culture and its consequences.

Hofstede defined culture as “the collective programming of the mind which distinguishes the members of one human group from another” (p. 25). He emphasizes that culture is shared among a human group, for example, a nation. Hofstede’s definition reflects an implicit view of culture, which takes a subjective approach and focuses on norms and values, beliefs and attitudes, and mentalities (Wuthnow & Witten, 1988), with an assumption that the implicit or informal aspects of the organization can be understood by unraveling the psychological processes that give meaning to surface manifestations of culture (Ouchi & Wikins, 1985). However, such an *entity view* of culture is incomplete and should be complemented with social and historical analysis (Kitayama, 2002).

Hofstede (1980a) identified four broad dimensions of national culture based on his extensive analysis of the survey data conducted in IBM from 1967 to 1973, which sampled about 11,600 people across 40 countries. These four dimensions are Collectivism-Individualism, Power Distance, Masculinity, and Uncertainty Avoidance. The first three dimensions have been consistently replicated in different studies using different survey instruments and samples (Hofstede & Bond, 1988). The fifth dimension, named Long-Term/Short-Term Orientation, was added based on results from studying Chinese values (Chinese Culture Connection, 1987). In this study, the fifth dimension will not be discussed as its research base is relatively weak and its implications for performance appraisal have been rarely discussed. Among the remaining four dimensions, only Individualism/Collectivism and Power Distance will be considered in this

study, for the following reasons: a) both dimensions concern values that are related to group and authority. Individualism/Collectivism and Power Distance affect the type of leadership most likely to be effective in a country (Hofstede & Bond,1988); b) the dimensions are the most intensively researched dimensions (Kirkman, Lowe & Gibson, 2006); c) the U.S. and China differ significantly in these two dimensions. For Collectivism/Individualism, the U.S. scores 91, ranking No. 1 among dozens of nations, while China scores 11, ranking one of the lowest. For Power Distance, the U.S. scores 38 and China scores 80, showing substantial differences. Hofstede (1980b) offered definitions for these two dimensions:

Collectivism: Characterized by the tight social framework in which people distinguish between in-groups and out-groups; they expect their in-group to look after them, and in exchange for that they feel they owe absolute loyalty to it

Individualism: A loosely knit social framework in which people are supposed to take care of themselves and of their immediate families only

Power Distance: The extent to which a society accepts the fact that power in institutions and organizations is distributed unequally

Culture and Self-Concept

Culture determines self-concept (Steers & Sanchez-Runde, 2002), and becomes part of self-concept through a “mutually constitutive relation” (Kitayama & Markus, 1999). Bandura (1999) also stressed the important role the cultural context plays in shaping self-concept, and stated, “A sense of selfhood is the

product of a complex interplay of social and personal construal process" (p. 230). Culture provides important contexts and frames of reference for individuals to grow. Individuals who interact with these contexts will develop stable thought and behavioral preferences that are unique to the contexts. Over time, such preferences "sink" into the unconsciousness world (Bateson, 1972) and constitute the implicit aspects of self-concept. As emphasized by Kitayama and Markus (1999), self-concept not only holds culture, it also reproduces culture; culture not only exists in self-concept, it also exists in collective behavior patterns. An easy tendency in culture research is to reduce culture to psychological processes and personal knowledge structures, and thus "psychologizing" it (Kitayama, 2002). To counter this tendency, it's necessary to take social and historical factors into consideration.

As contexts or practices are pluralistic and not necessarily cohesive, self-concept is also composed of loosely organized elements, which have different saliency in different circumstances (Bandura, 1999). For example, Markus and Kitayama (1991) showed that self-concept can be both independent and interdependent, and different selves dominate in different cultures. Triandis (1989) investigated how different aspects of the self were sampled in different social environments. In a more individualistic culture, the individual self is sampled more frequently than the collective self. In a more collectivistic culture, or when people are facing external threat, competition with out-groups, and common fate, the collective self is sampled more frequently. Cross-cultural studies of self-

concept have recognized many other differences between individualists and collectivists. For example, people in collectivist cultures pay more attention to context than content, while people in individualist cultures emphasize content over context (Miller, 1984; Cousins, 1989; Triandis, 2000). Thus, culture determines which aspect of the self or the environment is attended to. In turn, the self determines which needs are salient (Haslam, Powell & Turner, 2000). Striving for self-actualization and enhancing self-esteem become dominant needs when the individual self is salient, while the needs to enhance group-based self-esteem and to achieve group goals dominate when the collective self is salient. Therefore, culture can influence perception or behavior through self-concept and needs that tie to self-concept.

Chinese Culture and Workplace

Chinese culture is characterized by strong in-group collectivism. In studying cooperative behavior in different cultures, Chen, Chen, and Meidal (1998) claimed, “Group membership carries stronger psychological attachment for collectivists than for individualists. In Japan or China, for instance, such memberships as schools, geographical origins, or family names are more likely to trigger sufficient mutual identification to serve as a solid basis for emerging cooperation” (p. 293). Such group attachment can be traced back to familism highly valued by Confucius (Hofstede & Bond, 1988).

Around 500 B.C., Confucius, the most famous philosopher in ancient China, traveled across the nation to promote his philosophies. He proposed the

family should act as the prototype of the society and the order of society could be established by mirroring those mutually obligated and highly structured family relationships between father and son, elders and youngsters, and husband and wife. By specifying the roles and role relationships, Confucians were able to place individuals in a strict vertical structure. To make this structure work, individuals must be socialized into appropriate roles, and success was defined as the capability to attune to and meet the expectations of such roles. Confucians then promoted virtues that individuals should practice, such as compassion, loyalty, trust, righteousness, etiquette, kinship, moral integrity, filial piety, industriousness, thrift, perseverance, humility, tolerance, and empathy. People were advised to restrain their desires and needs, refrain from hedonism, and control their temper, in order to practice these virtues. Meanwhile, people were not encouraged to express their individualist selves in front of a group/family as the harmony of the group/family would be in jeopardy. Confucian teachings have strongly influenced Chinese culture for more than 2000 years.

The interdependent or the collective self will be constantly called on through role demands and expectations in a collectivist culture (Smith, 1983). Chinese people place great importance on anticipated reactions of others to their behavior (Yang, 1981). Their main achievement goals are honoring the clan and glorifying the ancestors (Yu & Yang, 1994). Here, the clan refers to an in-group that could consist of family, extended family, or extended relationships. Individual success is defined as being able to meet the expectation of the clan and promote

the clan's reputation. Being unable to fulfill the expectations may ruin the reputation of the clan and bring shame to the whole clan. In a Chinese expression, when a clan wins or loses, everyone in the clan will either earn or lose face, or *mien-tzu* (see Ho, 1976; Redding *&* Ng, 1983). The concern for *mien-tzu* exerts a mutually coercive power upon the members of a social network (Yau, 1986). *Mien-tzu*, or glories of a clan, can be accumulated, sustained, and reinforced through many ways, such as showing off luxury goods. Therefore, individuals from a collectivist culture have a strong obligation to meet other people's expectations and safeguard the reputation of the clan they are from. To fulfill external role demands, Chinese leaders must be more concerned with others' perceptions of themselves.

In China, institutions like teams, organizations and even the society have their roots in the family. Supervisors are supposed to display paternalism and safeguard all the interest of members, while subordinates will then reciprocate with compliance. It also makes intuitive sense that high collectivism should promote cooperation and teamwork as collectivists are more inclined to strive for super ordinate goals and place the group-identity ahead of the self-identity (Chen, Chen & Meindl, 1998; Kirkman, Lowe, & Gibson, 2006). However, such ideal situations are the exception rather than rule, as the Chinese workplace is far from unitary. In fact, Japan is the only nation that carries the collectivist culture over into its companies (Tayeb, 1994). Unlike Japanese organizations, Chinese organizations are characterized by clear boundaries between levels due to high

power distance. Employees form different in-groups which could either include members from the same level or across different levels. The workplace is thus compartmentalized into different cliques. Such cliques become an important form of social life, even out of the workplace. Dissimilar realities may be created because of intensive interaction and communication within a clique. The existence of multiple realities casts doubt on whether there exists a truth. “He (Confucius) dealt with Virtue, but left the question of Truth open” (Hofstede & Bond, 1989, p. 19). Chinese leaders have to maintain a subtle balance among multiple realities, which requires outstanding social skills (Hogan & Shelton, 1998).

Culture undergoes constant changes, especially when there are changes in labor mobility patterns and social movements (Earley & Gibson, 1998). The relationship between culture and the self is very dynamic (Kitayama & Markus, 1999). The self not only reproduces culture, it also changes it. Hofstede and Bond (1988) noticed that Japan, Taiwan, Hong Kong, and South Korea, which had enjoyed substantial economic growth from 1965 to 1985, placed more value on some Confucian virtues (e.g., thrift, persistence, shame, and authority/order), while devaluing some others (e.g., steadiness and stability, giving face, respect tradition, reciprocity). Ralston et al. (1999) compared three generations of Chinese Managers and found the new generation was becoming more individualistic and less collectivistic. The pace of such changes may accelerate due to globalization and new technology development.

Globally Dispersed Teams and Culture

Globalization and new technology development also makes working across nations a necessity. While expatriate assignment remains as an important form of international human resource management, more and more people across the world work in the same team and do not need to leave their home countries. Such work teams can be called “globally dispersed teams” (GDTs).

Many challenges exist for GDTs, such as differences in geography, language, and culture, among which cultural barriers could be the most difficult to overcome. People from different cultures have different beliefs and see things through different frames of reference, easily resulting in miscommunication, conflict, and even distrust (Baba, Gluesing, Ratner, & Wagner, 2004). Usually, it is rare for people to bring the cultural context into awareness and discuss it openly (Cramton, 2001). Leaders of GDTs must take the responsibility, work through different cultures, and promote cultural integration in the team (Black & Mendenhall, 1990). Many different skills or personal characteristics that can help people better perform in foreign cultures have been identified, such as flexibility/adaptability, relational skills, interpersonal skills, extra-cultural openness, job knowledge and motivation, perceptual skills, and self-efficacy (Arthur & Bennett, 1995; Black, Mendenhall, & Oddou, 1991; Mendenhall & Oddou, 1985). Equipped with these skills and characteristics, people can communicate better and improve performance accordingly.

Summary for This Section

This section shares Kitayama's (2002) *system view* of culture, which states that culture exists both *in the head* and *out there*. It is suggested that there exists a reciprocal relationship between culture and self-concept. While culture plays a critical role in shaping self-concept, self-concept also reproduces and transforms culture (Kitayama & Markus, 1999). The U.S. and China have very different cultures, with the former being high on individualism and low on power distance and the latter being high on collectivism and high on power distance. A historical description of the Chinese culture further shows that it features familism, status, Confucian virtues, in-group, and *mien-tzu*. The Chinese workplace is usually compartmentalized into different cliques, resulting from in-group practices and status differences. Globally dispersed teams (GDTs) are characterized with team members from different cultures working together in different nations. Cultural diversity poses enormous challenges to GDT leaders, because they have to work across cultural boundaries. People from different cultural backgrounds may have different perceptions of their leadership effectiveness. Self-other agreement becomes harder to achieve. The next section will discuss the relationship between culture and self-other agreement.

Culture and Self-Other Agreement

With a belief that culture will affect self-other agreement, Atwater et al. (2005) compared self-other agreement between the U.S. and several European nations and found that the relationship between self-other agreement and

performance differed across cultures. An important finding is that French differentiated high performers less well than their U.S. counterparts, possibly due to the leveling effects unique in the French culture. With a more feminine culture, French do not try to be better than others. Atwater et al. urged research in other cultural contexts with higher collectivism scores, believing that self-other agreement would be more strongly influenced by Individualism/Collectivism.

Culture can affect self-other agreement in two ways. A more direct way is through self-concept or meaning systems. A rater and a ratee from two cultures may have different schemas of effective leadership behavior and thus perceive the same behavior differently. For example, in a globally dispersed team (GDT), a supervisor from an individualist culture may believe autonomy is important and subordinates should make their own decisions, but subordinates from a collectivist culture may perceive the supervisor's behaviors less plausible as he/she is supposed to take the responsibility to make the decision for them. Subordinates may see the supervisor as indecisive. Such different schemas could cause self-other discrepancies. The other, less direct way is through collective behavioral patterns (Kitayama & Markus, 1999; Kitayama, 2002) that are unique to each culture. For example, a supervisor in a high power distance culture may hesitate to solicit feedback from subordinates (Smith et al., 1994), which in turn results in self-other discrepancies. Kenny's (1991) weighted-average model provides a convenient framework to investigate cultural effects through collective behavioral patterns.

The U.S. and China

This study attempts to understand the relationship between self-other agreement and performance in the U.S. and China. Within each culture, the self-concept will be more similar than different. Hence, discussion will focus on collective behavior patterns. Table 3 lists culture's possible impact on Kenny's (1991) factors. In the table, a high collectivism/high power distance (H/H) culture is contrasted to a low collectivism/low power distance (L/L) culture. The former is represented by China while the latter corresponds to the U.S. A plus sign represents a culture that may promote a factor; a minus sign represents a culture that may constrain a factor; "mixed" represents a culture that either promotes or constrains a factor.

Acquaintance refers to the amount of information the rater has about the ratee. Kenny's (1991) projected that acquaintance will have a large effect on consensus when people get to know each other but the effect levels off pretty quickly. Considering a one-year timeframe, acquaintance will have a relatively small impact on self-other agreement. High collectivism is associated with more positive attitudes toward teams (Kirkman et al., 2006), but high power distance keeps the leader far from subordinates hierarchically speaking (Bochner & Hesketh, 1994). On the contrary, individualists prefer to work alone (Earley, 1993; Jung & Avolio, 1999), but low power distance creates fewer barriers for them to interact with their supervisors. Overall, raters from both cultures should have adequate behavioral information to rate the ratees.

Table 3

Disparate Cultural Influences on Kenny's (1991) Factors

Kenny Factors	Individualism/ Low Power Distance	Collectivism/ High Power Distance
Acquaintance	Mixed	Mixed
Behavioral Consistency	+	-
Overlap	+	-
Similar Meaning Systems	+	-
Communication	+	-
Unique Impression	-	+

It has been speculated that the individual identity is salient in an individualist culture, while in a collectivist culture the collective identity is salient (Lord & Brown, 2004). When the individual identity dominates, a person will display more behavioral consistency, but less will be expected for collectivists (Cousins,1984; Kitayama & Markus, 1999). In a H/H culture, people tend to demonstrate different behavioral patterns according to status and relationship patterns. In-group practices and status differences compartmentalize the workplace so that the rater can only observe limited behaviors of the ratee. The ratee is also supposed to adjust their behaviors to different raters and situations.

Therefore, raters are less likely to observe the same behavior of a ratee in a H/H culture in comparison to those in a L/L culture.

In-group practices and status differences make leadership behavior less observable and less consistent, and they also affect how the observed behaviors are perceived by different raters. The fundamental assumption of MSF (Murphy & Cleveland, 1995) is that different meaning systems exist across levels in the same culture. However, support for this statement is generally weak (e.g., Mount et al., 1998; Scullen et al., 2000; Yammarino, 2003). Raters from different levels seem to have similar understandings of effective performance. Nevertheless, it should be noted that most studies cited are based on U.S. participants, who come from a L/L culture. It is possible that in a L/L culture people are more inclined to use the same external standards for rating than those in a H/H culture. Also, consistent expectations or standards are better communicated in a low power distance culture. In a H/H culture, however, different realities may exist because in-group practices and status differences tend to compartmentalize the workplace into cliques. Different realities correspond to different meaning systems.

Communication makes consensus more likely (Kenny, 1991). In a L/L culture, people communicate expectations and seek feedback directly (Sully De Lique & Sommer, 2000). Direct communication helps to increase self-other agreement. In contrast, people tend not to communicate expectations and feedback explicitly in a H/H culture. Bochner and Hesketh (1994) noted that in a

high power distance culture subordinates are less open with their superiors. Seeking feedback could bear face costs and people use monitoring and indirect inquiry feedback-seeking strategies (Sully De Lique & Sommer, 2000). In the same vein, giving others feedback could be risky as others may feel embarrassed, especially when the feedback is conducted across levels and between cliques. With constrained communication between the rater and the ratee, self-other agreement will be more difficult to achieve in a H/H culture. But when both raters are from the same clique, consensus is much more likely due to intensified communication in a clique.

Unique Impression refers to the degree to which the rater uses information other than observed behaviors. Table 2 suggests that a H/H culture would incur more unique impression. High power distance makes behavior less observable so people in a H/H culture are more likely to go to their in-group for information. As a result, the rater bases his/her ratings on perceptions learned from others instead of behaviors they actually observe. In addition, leaders in a collectivist culture usually have more informal contact with fellow workers than those in an individualist culture (Bochner & Hesketh, 1994). Due to in-group practices and status differences in the Chinese workplace, leaders may develop disparate personal relationships with their subordinates. Such informal contact or personal relationship can introduce unique impression to the rating process. In contrast people from a L/L culture will rely more on independent judgment, with less influence from others.

Globally Dispersed Teams

Unlike the U.S. or China, globally dispersed teams (GDTs) are characterized by heterogeneous cultures, which may exert a negative influence on self-other agreement. Because different cultures provide different lenses or views of the world, people with different origins may perceive the same behavior of a leader differently. As a result, self-other agreement will be more difficult to achieve. The more cultures a team has and the more different these cultures are, the lower self-other agreement can be.

With multiple cultures in a team, collective behavior patterns will be less consistent and their impact on self-other agreement will be less systematic. For example, a manager from an individualist culture may demonstrate high behavioral consistency to his/her team, while another manager from a collectivist culture may attempt to meet others' expectations by displaying different behaviors. Thus, behavioral consistency may vary in different GDTs. Culture's impact on acquaintance, overlap, communication, and unique impression is similarly equivocal. To better understand the effects of these factors on self-other agreement and consensus between raters, there is a need to take geographical, language, and technological elements into consideration. Geographical and language differences suggest less communication and less unique impression while use of standard communication processes and technologies suggests increased overlap.

Summary for This Section

In sum, this section introduces two ways in which culture can influence self-other agreement. First, people from different cultures will experience more difficulty to reach consensus on ratings, due to different schemas they have. Second, culture influences self-other agreement through collective behavioral patterns. Table 2 indicates that raters in a H/H culture have particular difficulty in giving ratings that are in agreement with the ratee, due to lower behavioral consistency, less overlap, compartmentalized meaning systems, restricted communication, and more unique impression. As a result, self-other agreement will be harder to achieve in a H/H culture than in a L/L culture. In GDTs, the existence of multiple meaning systems makes achieving self-other agreement more difficult. However, behavioral consistency, overlap, communication, and unique impression are less impacted by culture than geographical, language, and technological factors. Based on these new insights, several research hypotheses will be presented in the next section.

Research Question and Hypotheses

MSF has attracted more and more research attention during the past 15 years but practitioners and researchers are still perplexed by the rating discrepancies between different sources, especially, self-other discrepancies. Recent research has shown that such discrepancies are stable and they predict actual performance. Overestimation corresponds to lower performance while ratings in agreement predict higher performance. However, little work exists

regarding how the rater and the ratee come to agreement on their ratings. Moreover, as MSF expands into the global arena, self-other agreement becomes a bigger concern. So far, little is known about culture's role in self-other agreement. This study aims to develop a conceptual framework that describes how the rater and the ratee reach agreement and how contextual factors affect the process. Heavily borrowed from person perception literature, a conceptual framework is developed based on Hanges et al.'s (2000) connectionist model of leadership perception and Kenny's (1991) weighted-average model. Culture is built in as an important contextual factor that affects self-other agreement. This study also compares samples from the U.S., China, and GDTs to investigate how self-other agreement is different in different cultural contexts.

The U.S. and China

It has been suggested that Chinese leaders are more concerned with others' perceptions of themselves in order to meet other people's expectations. However, it's also suggested that the Chinese workplace is more than unitary. It is often compartmentalized into multiple realities. Chinese leaders need to strike a delicate balance among these realities and manage a coherent outlook to each. Success often depends on whether the collective self has been transcended beyond the clique and whether commitment to superordinate goals and the group identity has been invoked (Chen et al., 1998). Charismatic and participative leadership is highly valued in a collectivist culture (Kirkman, et al., 2006), as it helps to promote social identify and satisfy employees' need for group-based

self-esteem (Haslam et al., 2000). Good leaders will be those who can respond resourcefully to various role demands and gain unanimous commitment from different constituents.

Self-other agreement is more likely for people who care about others' perceptions than for those who are self-focused. The collective self is more loosely organized and less independent than the individual self (Markus & Kitayama, 1991). As a result, people from a collectivist culture are less likely to derive ratings directly from the self-concept than those from an individualist culture. That is, people from a collectivist culture may focus more on behaviors when they conduct self-ratings. With more emphasis on behavior than the self-concept, self-other agreement could improve. However, collectivists are less likely to communicate with others about their performance, which might decrease self-other agreement. In addition, because collectivists tend to demonstrate different behaviors in different contexts, they may at times have problem in rating their own behaviors accurately, resulting in lower self-other agreement. Overall, I propose:

H1: The correlations between self-ratings and other-ratings (supervisors, subordinates, and peers) will be lower in the U.S. than those in China.

In contrast, a compartmentalized organizational reality may make consensus between raters more difficult to achieve. The ratee may demonstrate different behaviors in different situations and raters may observe a small set of the ratee's behaviors. Moreover, multiple meaning systems may exist and raters

may form unique impression through informal interactions with the ratee. Finally, communication is usually restricted between cliques. As a result, rating consensus between others will be lower in a H/H culture than in a L/L culture. Performance will become less critical in determining ratings when multiple realities exist. Therefore, I propose:

H2: The correlations between other-ratings, including supervisor-ratings of performance, will be higher in the U.S. than those in China.

Overestimation has been found to predict poorer performance in the L/L culture (Atwater & Yammarino, 1997; Van Veslor et al., 1993; Yammarino & Atwater, 1993). Seeing oneself better than how others see him/her could be especially problematic in a H/H culture. A major challenge for leaders from a H/H culture is to meet the expectations of different cliques. As the expectations are often inconsistent, and sometimes even in conflict, leaders need to manage them artfully so that harmony can be achieved. Failing to do so may result in backfiring from those cliques whose expectations are not met. Overestimation may indicate that the ratee pays little attention to the expectations of others, lacks social skills to present others a coherent impression, is overly contented with himself/herself, or is discounting negative feedback. Over-estimators will have difficulty to meet intensive role demands and earn others' support. As result, their performance will greatly suffer. Therefore, I propose:

H3: Overestimation (self-ratings > other-ratings) will be associated with poorer supervisor-rating of performance in China as compared with the U.S.

In contrast, the relationship between underestimation and performance is mixed (Atwater & Yammarino, 1997; Yammarino & Atwater, 1993). Underestimation may reflect that the ratee has low self-worth (Yammarino & Atwater, 1997). However, In a H/H culture, underestimation is also associated with such Confucian virtues as hardworking, perseverance, humility, tolerance, and empathy. Underestimation may indicate the ratee can effectively respond to other's expectations, although may not necessarily be able to meet various role demands and gain others' support. Overall, I propose:

H4: Underestimation (self-ratings < other-ratings) will be associated with poorer managerial performance in the U.S. as compared with China.

Globally Dispersed Teams

Another major concern for MSF is about its effectiveness in globally dispersed teams (GDTs), where people are working together across national boundaries. Working in multiple locations and with multiple cultures makes performance appraisal difficult for teams to carry out. MSF is a promising alternative as it take multiple perspectives into consideration. Based on the framework developed in this study, it is possible to investigate how geographical and cultural factors influence the MSF process. Research on MSF application in GDTs can not only show us whether MSF is applicable in another setting, but also further our understanding of culture's role in self-other agreement.

As GDTs are often constituted by people from different cultures, members usually have different meaning systems. Dissimilar meaning systems makes

reaching self-other agreement particularly difficult, especially when there are additional language and geographical barriers. Therefore, I propose:

H5: The correlations between self-ratings and other-ratings (supervisors, subordinates, and peers) will be higher in the U.S. than those in GDTs.

It may be counterintuitive but geographical and technological factors may actually help to increase consensus between raters, despite dissimilar meaning systems. GDTs use more virtual communication tools such as emails and conference calls, which usually are highly structured for the sake of efficiency and effectiveness. As demonstrated above in the discussion of assessment centers, consensus between raters can improve in a more structured situation, where contextual factors are under control. Unlike in physical workplaces, where individual face-to-face interactions are the norm, in virtual situations people often receive the same emails or attend the same conference calls. Overlap can be actually higher in this situation than traditional face-to-face situations. In addition, using technologies may also constrain how information is delivered, which could actually result in more consistency in terms of a leader's behavior. Finally, without intense personal connections or interactions, unique impression will be lower, meaning that raters will focus more on behavior and performance when they do ratings. Therefore, I propose:

H6: The correlations between other-ratings, including supervisor-ratings of performance, will be similar in the U.S. and GDTs.

In globally dispersed teams, a ratee usually has fewer opportunities to observe other's reactions to his/her behavior face-to-face and communicate with others about his/her performance, so it will take extra effort to achieve self-other agreement. Overestimation may reflect a ratee's inability to understand his own behavior in another cultural setting or communicate about her own performance across cultures. As a result, ratings may be heavily influenced by the self-concept which is overly optimistic. Being open and non-judgmental is important in working with people from other cultures (Mendenhall & Oddou, 1985). A tendency to be self-contented and cut oneself off from feedback may result in poor performance. Therefore, I propose:

H7: Self-other overestimation (self-ratings > other-ratings) will be associated with poorer managerial performance in globally dispersed teams as compared with the U.S.

In contrast, under-estimators tend to be sensitive to cultural differences and take them seriously. They will try to understand the effects of their behaviors in another cultural setting. Once there is an issue, they will put in extra effort to address it. However, underestimation may also reflect low self-efficacy, which may reflect poor adjustment to different cultures (Black, Mendenhall, & Oddou, 1991). According to Black et al., poor adjustment to different cultures may be exacerbated when feedback is in short supply. Therefore, I propose:

H8: Self-other underestimation (self-ratings < other-ratings) will be associated with poorer managerial performance in globally dispersed teams when compared to teams in the U.S.

CHAPTER III

METHOD

Participants

Data were collected from managers in the U.S., China, and globally dispersed teams (GDTs) who participated in a leadership development program. As part of each program, the manager, his/her supervisor(s), peers, and subordinates completed a multi-rater feedback instrument. A total of 1000 managers who identified themselves as non-expatriates and their work country as the U.S. were randomly sampled from a data base with more than 27,000 managers. Those with both self-rating and supervisor-rating scores missing were removed from the original sample, leaving the final sample size of the U.S. sample to 818. A total of 428 managers who were from mainland China, Hong Kong, and Taiwan formed another sample, with those who had neither subordinate-rating scores nor peer rating scores being excluded. Finally, a total of 871 managers who were part of globally dispersed teams formed the GDT sample. To be included in the GDT sample, a manager must have had at least 40% of his/her raters from nations other than his/her working nation. For subordinate- and peer-ratings, at least two respondents were required; otherwise ratings were treated as missing. In the Chinese sample, the number of raters ranged from 4 to 26, with a median of 9. About 18% of managers had only one or

no subordinates rating them, and about 8% of them had one or no peers rating them. In the U.S. sample, the number of raters ranged from 2 to 23, with a median of 10. About 13% of managers had one or no subordinates rating them; about 6% of them had one or no peers rating them. For the GDT sample, the number of raters ranged from 2 to 30, with a median of 11. About 8% of managers had one or no subordinates rating them, and about 4% of them had one or no peers rating them. For all three samples, it was typical that a manager had one supervisor rater, 3 subordinate raters, and 4 peer raters.

In the U.S. sample, 66% of the managers were male, the average age was 44 years, 87% were white, 80% had bachelor degrees or above, 60% were at the level of middle management or higher. In the Chinese sample, 66% were male, the average age was 41 years, 86% had bachelor degrees or above, 53% were at the level of middle management or higher. In the GDT sample, 77% were male, the average age was 43 years, 87% had bachelor degrees or above, 68% were at the level of middle management or higher. Based on available data, all three samples covered a wide range of industries, with relatively more participants from Banking and Finance, Electrical/Electronics, Retail Trade, and Light Manufacturing. The GDT sample consisted of more participants from the banking and finance industry than the other two samples. A wide range of functions were covered in all three samples. Relatively more participants were from Sales, Finance/Accounting, Marketing, and Operations. Overall, these three samples were comparable in terms of industry and function distributions.

Measures

In this study, both independent and dependent variables were derived from the Profilor, a multirater feedback instrument developed by Personnel Decisions International Corporation (Hezlett, Ronnkvist, Holt, & Hazucha, 1997) to collect competency/performance ratings. The independent variables were self- and other-ratings of leadership and the dependent variable was supervisor-ratings of overall performance.

The Profilor

The Profilor contained 135 items, grouped by the publisher into 24 scales, i.e. 24 competencies. These competencies drew heavily on the Campbell, Dunnette, Lawler and Weick (1970) research but were updated in 1990 to reflect the changing nature of managerial work. The development of the Profilor was a rigorous process. It is based on several decades of consulting experience and research on management, including an extensive review of the management and psychology literatures, exhaustive analysis of the large Management Skills Profile (MSP) data base that included more than 20,000 managers, thorough job analysis, group interviews with hundreds of managers representing many functional areas and most major industries, and pilot testing. The Profilor was intended to represent behavioral performance competencies that were generally required of managers in a changing business environment. The instrument was designed to identify interactive transformational leadership which paid special attention to team and participative leadership.

Respondents were told that the surveys were for gathering feedback information for the manager's development only and confidentiality was assured. First, respondents (except supervisors) were anonymous. Second, ratings were aggregated for each perspective, so no individual responses could be identified. Third, only the manager being rated would see the aggregated results. The process of filling out the feedback survey was quite simple. Respondents would receive an electronic invitation from the focal manager and they just needed to get online and complete the survey step by step as instructed.

All surveys were in English. Most managers and respondents from China and globally dispersed teams worked in multinational companies and English was their working language. The survey adopted a 5-point scale that indicated the frequency with which the manager engaged in each behavior. The response scale ranged from 1 (not at all) to 5 (to a very great extent). When respondents didn't feel they had enough opportunity to observe a certain behavior, they could choose "Not Applicable". Items were short behavioral descriptions, such as "Clearly expresses ideas and concepts in writing" and "Recognizes and utilizes the contributions of people from diverse backgrounds." There were three additional open-ended questions concerning what the focal manager had done effectively, what new skills or behaviors could make him/her more effective, and what other suggestions could promote his/her insight and development.

The Profilor had adequate reliability and validity. Each competency consisted of 4-9 items, with internal consistency reliability (Cronbach's alphas) for

self-rating ranging from .65 to .83, supervisor-rating from .75 to .89, subordinate-rating from .85 to .94, and peer-rating from .82 to .93. Higher reliability existed for subordinate-rating and peer-rating because usually more subordinates and peers responded to the survey than supervisors. Another reliability index, interrater reliability, could be more relevant in MSF as the main concern was whether different raters agreed on a certain behavior in terms of rating. Intraclass correlation coefficients based on three raters ranged from .47 to .60 for peers and from .48 to .61 for subordinates, lending further support for the reliability of the Profilor. Intraclass correlation coefficients for supervisors were not available as in very few cases the number of supervisors would go over 2.

Some evidence of validity was also offered. Ratings correlated from .55 to .84 with overall performance which was based on a five-item performance measure. Ratings were also found to be related to competence and long-range potential/promotability, which were measured in the beginning of the survey. The competence measure was assessed by one item: “How would you rate this manager’s competence in his/her current position?” It was evaluated on a 7-point scale ranging from 1 (“Outstanding; one of the best”) to 7 (“Very weak; one of the worst”). The potential measure was also assessed by one item: “How would you rate this manager’s long-term potential in his/her career?” It was evaluated on multiple levels ranging from Senior Executive to First-Line Management. The correlations between ratings and competence and potential were moderate to high.

Leadership Ratings

In this study, only seven competencies that assessed leadership were used, which was consistent with the Atwater et al. (2005) study. These seven competencies were related to directing, influencing, motivating, leading courageously, fostering teamwork, coaching and developing, and championing change. Factor analyses showed that there was one factor underlying these competencies, regardless of the sample source or the perspective. These seven competencies were added up to form an overall leadership measure. The internal consistency coefficients (Cronbach's alpha) of the overall leadership measure for each perspective across samples ranged from .93 to .98, indicating that the new measure was highly reliable.

Performance Ratings

Supervisor-ratings of a five-item performance measure were used as an outcome measure. The items were: "gets the job done", "is an effective manager overall," "get the work done on time," "produces high quality work," and "accomplishes a great deal." These items were rated on a 5-point scale. Using the supervisor-rating as a criterion was reasonable as it's the most reliable and valid measure among ratings from different sources (Conway & Huffcutt, 1997). The internal consistency reliability coefficients (Cronbach's alphas) for the U.S., the Chinese, and the GDT samples were .88, .87, and .85, respectively.

Data Analysis

Two procedures were used to test the hypotheses. An EQS procedure was used to test the equality of the covariance matrices derived from two different samples. A polynomial regression procedure was used to explore the relationship between self-other agreement and performance in each culture. But before applying the polynomial regression procedure, a multivariate regression procedure that used self-other ratings as dependent variables (Edwards, 1995) was adopted to investigate whether demographic variables played a role in self-other agreement.

The EQS Procedure

Instead of testing correlation coefficients across samples one by one, the EQS procedure was able to test the correlation coefficients in two samples simultaneously by constraining all or some of the covariances to equal. The EQS procedure would compare the observed covariances to those predicted in the specified model. It provided many fit indices that could indicate whether the model specified (with all covariance being constrained to equal) was accepted or refuted. These indices included chi-square, Goodness-of-Fit Index (GFI), Adjusted Goodness-of-Fit Index (AGFI), Comparative Fit Index (CFI), Standardized Root Mean Square Residuals (RMSR), Root Mean Square Error of Approximation (RMSEA), and so on. The chi-square value reflected the discrepancy between the modeled covariance matrix and the observed covariance matrix. A good model fit required the chi-square value to reach non-

significance, but a model might still be accepted based on other indices despite a significant chi-square value, due to its sensitivity to sample size. GFI was a classic fit index that represented how much the observed covariances were accounted for by the covariances implied by the model. GFI could be inflated in the condition of large sample size but adjusted GFI (AGFI) was less influenced. CFI was more complicated than GFI and AGFI in that it also compared the observed covariances and modeled covariances to a null model where all covariances were zeros. CFI was least affected by sample size among various fit indices. SRMR referred to the average difference between the observed and modeled variances and covariances, based on standardized residuals. RMSEA was a commonly used index based on chi-square and the model's degree of freedom. It represented discrepancy per degree of freedom. For CFI, GFI, and AGFI, a value larger than .90 indicated an adequate fit. For RMR and RMSEA, values lower than .08 and .06 indicated a good fit, respectively.

The Polynomial Regression Procedure

Different indices (e.g., algebraic difference, absolute difference, squared difference, and sum of absolute differences) had been suggested to study difference scores but their use was criticized for imposing untenable constraints (e.g., Edwards, 1994). An often used algebraic difference equation was:

$$Z = b_0 + b_1(X - Y) + e, \quad (1)$$

where X and Y were predictors of Z, b_0 represented the intercept, b_1 represented the slope, while e represented the random measurement error. The equation could be expanded into:

$$Z = b_0 + b_1X – b_1Y + e. \quad (2)$$

Here, researchers made an untenable assumption that the slopes for X and Y were equal in magnitude but opposite in sign. A more appropriate equation would be:

$$Z = b_0 + b_1X + b_2Y + e. \quad (3)$$

Recognizing such a limitation, Edwards (1994) suggested using polynomial regression. The use of this procedure was based on following assumptions: (a) the relationship between two different scores and an outcome should be considered in three dimensions, (b) Three-dimensional response surfaces could be used to depict such a relationship, (c) The plausibility of the constraints could be tested using a hypothesis testing approach (Kristof, 1996). In this study, the polynomial regression procedure would be used to investigate the relationship between self-other agreement and performance. To account for possible curvilinear relationships, instead of using:

$$Z = b_0 + b_1(X – Y) + b_2(X –Y)^2 + e, \quad (4)$$

or the expanded one:

$$Z = b_0 + b_1X – b_1Y + b_2X^2 - 2b_2XY + b_2Y^2 + e, \quad (5)$$

a full quadratic equation would be adopted:

$$Z = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2 + e. \quad (6)$$

The polynomial regression procedure would be conducted with three samples combined, instead of running it on each sample separately. This procedure had two advantages. First, a larger sample size would help to reduce standard errors. Second, differences between parameters in two samples could be identified effortlessly. Two dummy variables would be created to represent the three samples. These dummy variables were added to Equation 6 as moderators. Significant moderating effects would indicate significantly different regression parameters between two samples.

A response surface not only provided vividness but also brought complexity. Fortunately, Edwards and Parry (1993) offered some convenient tools to facilitate interpretation. A three-dimensional figure usually took many different forms, such as plane, convex, concave, and saddle. For these figures, the key was to find out the slope and the curvature in some interesting areas. The first area was the stationary point, where the slope of the surface was zero in all directions. For some surface forms, such as plane, no stationary point existed. The second interesting area was the principal axes of the surface (the first principle axe and the second principle axe), which ran across the stationary point perpendicularly. These axes represented either the most or the least curvature, depending on the form of the surface. The third interesting area was along the X = Y line and the X = -Y line, both being related to self-other agreement. According to Edwards and Parry (1993), when X = Y, the full quadratic equation could be transformed into:

$$Z = b_0 + (b_1 + b_{2)}X + (b_3 + b_4 + b_5) X^2 + e, \quad (7)$$

in which $(b_1 + b_{2)}$ represented the slope at (0, 0), i.e., when both X and Y were 0, and the quantity $(b_3 + b_4 + b_5)$ represented the curvature along the line. When X= -Y was interested, the full quadratic equation could be transformed into

$$Z = b_0 + (b_1 - b_{2)}X + (b_3 - b_4 + b_5) X^2 + e. \quad (8)$$

By identifying the stationary point, the principle axes, and the slope and the curvature along the X = Y (or X = -Y) line, the response surface could be clearly interpreted.

Data Analysis Strategies

To test hypotheses H1, H2, H5, and H6, the covariance matrices of three samples were created and EQS procedures were run to compare these covariance matrices by pairs. By specifying different model, i.e., loosing some constraints, covariances of different sizes could be identified. To test Hypotheses H3, H4, H7, and H8, a polynomial regression procedure was applied with sample source as controls. Supervisor-ratings of performance were regressed on self-ratings and other-ratings in the first step (main effects), and the product of self-ratings times other-ratings, the square of self-ratings and square of other-ratings in the second step (higher order effects). Self-, subordinate-, and peer-ratings, were also centered at the same value, based on the midpoint of their shared scale, in order to reduce multicollinearity (Kristof, 1996).

CHAPTER FOUR

RESULTS

Descriptive Statistics

Table 4 shows the means, standard deviations, and correlations coefficients among different ratings across three samples. The means in the global dispersed team (GDT) sample were close to those in the U.S. sample, but the means in the Chinese sample were consistently lower. With only competency ratings considered, the means of self-ratings were the closest to subordinate-ratings, followed by those of peer- and supervisor-ratings, with peer-ratings averaging the lowest. Supervisors and peers in the Chinese sample tended to give lower ratings than did their GDT and U.S. counterparts. The standard deviation was the lowest for peer-ratings in the Chinese sample (*SD* = .366) and the highest for supervisor-ratings of performance in the U.S. sample (*SD* = .600). Supervisor-ratings had the largest standard deviations while variations in peers were the smallest. Self-ratings in the Chinese sample displayed more variation than those in the U.S. sample. In contrast, the standard deviations of supervisor-ratings in the U.S. sample were larger than those in the Chinese sample.

The pattern of inter-correlations was similar in three samples: the correlations between self- and other-ratings were generally smaller than those between other-ratings. Self-ratings correlated the most strongly with subordinate-

ratings and the most weakly with performance. The correlation between self- and other-ratings was the highest in the Chinese sample and the lowest in the GDT sample. Among the correlations between other-ratings, the correlation between peer-ratings and supervisor-ratings of leadership was higher in the U.S. and the GDT samples, and the one in the Chinese sample was substantially lower. Instead, the correlation between subordinate- and peer-ratings was the highest in the Chinese sample. The correlations between supervisor-ratings of leadership and performance were consistently in mid .70s, indicating that these two measures shared a large amount of common variance. Supervisor-ratings of performance appeared to have lower correlations with other-ratings than did supervisor-ratings of leadership.

Table 4

Descriptive Statistics and Correlations by Sample

				Variables			
Persepective	*n*	*M*	*SD*	Self	Sub	Peer	Sup
All Participants							
Self	2117	3.767	.421				
Subordinate	1862	3.782	.450	.250**			
Peer	1995	3.654	.377	.213**	.382**		
Supervisor	2117	3.692	.484	.201**	.329**	.424**	

Performance	2117	4.012	.581	.112**	.279**	.340**	.764**
The U.S.							
Self	818	3.782	.405				
Subordinate	715	3.830	.458	.285**			
Peer	766	3.695	.377	.205**	.349**		
Supervisor	818	3.762	.509	.195**	.351**	.430**	
Performance	818	4.094	.600	.125**	.289**	.333**	.751**
China							
Self	428	3.700	.441				
Subordinate	349	3.713	.448	.304**			
Peer	395	3.535	.366	.296**	.422**		
Supervisor	428	3.529	.449	.232**	.284**	.303**	
Performance	428	3.845	.536	.160**	.270**	.221**	.751**
GDTs							
Self	871	3.784	.422				
Subordinate	798	3.768	.440	.191**			
Peer	834	3.672	.372	.159**	.383**		
Supervisor	871	3.707	.458	.169**	.306**	.434**	
Performance	871	4.018	.566	.054	.252**	.361**	.768**

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed).

EQS Results

Three EQS statistical procedures were run to test the equality of covariance matrices for three pairs of samples (U.S. and China, U.S. and GDTs, China and GDTs), respectively. Table 5 lists different models and their fit indices. Model 1 represented the model with all covariances being constrained to equal and Model 2 represented the model with some untenable constraints being loosened. A constraint was untenable only when the multivariate Lagrange Multiplier (LM) test in EQS suggested removing a constraint would improve the model fit significantly. With some constraints being removed, the fit indices in Model 2 would show some improvements, featured with higher CPI, GPI, and AGPI, as well as lower Chi Square (λ^2), SRMR, and RMSEA.

U.S. versus China

The model which specified that the U.S. sample and the Chinese sample shared the same covariance matrix was generally supported, with $\lambda^2(10) = 10.860$, $p = n.s.$. The indices of CPI, GPI, and AGPI were all larger than .95, indicating a good fit of the model. The indices of SRMR and RMSEA were smaller than .08 and .06, respectively, also indicating a good fit. However, the LM test revealed that loosening the constraint between peer-ratings and supervisor-ratings of leadership would improve the fit of the model further. The fit indices of the new model suggested a better fit, with $\lambda^2(9) = 7.998$, $p = .n.s.$. The CPI index reached 1.000, indicating a perfect fit. The indices of GFI and AGFI increased slightly, while the values of SRMR and RMSEA became even lower, showing

some improvements over the original model. Hypothesis 1 stated that self-ratings would correlate higher with other-ratings in the Chinese sample than in the U.S. sample. Because the U.S. sample and the Chinese sample had very similar covariance matrices, Hypothesis 1 was not supported. Hypothesis 2 claimed that consensus between other-ratings would be lower for Chinese. Peer-ratings were found to correlate lower with supervisor-ratings of leadership in the Chinese sample than in the U.S. sample. Therefore, Hypothesis 2 was partially supported.

Table 5

Fit Indices for the Equality of Covariance Tests

	Indices					
	λ^2	CFI	GFI	AGFI	SRMR	RMSEA
China vs. U.S.						
Model1 (*df* = 10)	10.86	.999	.996	.987	.028	.009
Model2 (*df* = 9)	7.998	1.000	.997	.989	.023	.000
U.S. vs. GDTs						
Model1 (*df* = 10)	9.762	1.000	.997	.992	.024	.000
China vs. GDTs						
Model1 (*df* = 10)	17.924	.995	.993	.980	.048	.027
Model 2(*df* = 6)	7.824	.999	.997	.986	.030	.017

U.S. versus GDTs

A similar procedure was conducted between the GDT and the U.S. samples. The results demonstrated that the covariance matrix of the GDT sample was equivalent to that of the U.S. sample, with $\lambda^2(10) = 9.762$, $p = n.s.$. The values of CPI, GFI, and AGFI were all larger than .95, while SRMR and RMSEA remained quite small. Specifically, the index of CPI equaled to 1.000, indicating a perfect fit of the model. The indices of GPI and AGPI were above .99, indicating a nearly perfect fit of the model. The value of SRMR was .024, which was substantially lower than .08, while the value of RMSEA approached zero, also indicating a perfect fit of the model. Hypothesis 5 claimed that self-other agreement would be lower in the GDT sample than the U.S. sample. The EQS results failed to support the hypothesis. However, Hypothesis 6 was supported, as the correlations between other-ratings in the GDT sample remained close to those in the U.S. sample.

China versus GDTs

Since the self-other correlations were the highest in the Chinese sample and the lowest in the GDT sample, it would be interesting to see whether these two covariance matrices equaled to each other, so an additional EQS statistical procedure was conducted. The model with all covariances being constrained equal was only marginally supported, with $\lambda^2(10) = 17.924$, $p = .056$. After loosening the constraints between self-ratings and subordinate-ratings, between self-ratings and peer-ratings, between self-ratings and supervisor-ratings of

performance, and between peer-ratings and supervisor-ratings of leadership, the fit of the model improved greatly, with $\lambda^2(6) = 7.824$, $p = n.s.$. CPI increased from .995 to .999, GPI increased from .993 to .997, and AGPI increased from .980 to .986. In contrast, the value of SRMR decreased from .048 to .030 and that of RMSEA reduced to .017 from .027, showing some substantial improvements over the original model. These results suggested that self-other agreement was easier to achieve in the Chinese sample than in the GDT sample. Also, consensus between peer-ratings and supervisor-ratings of leadership was higher in the GDT sample than in the Chinese sample, consistent with what was found in the previous comparison between the Chinese and the U.S. samples.

Demographics and Self-Other Agreement

Before moving on to investigate possible relationships between self-other agreement and performance, a multivariate regression procedure was used to explore the effects of demographic factors on self-other agreement. Because the results of demographic analyses demonstrated some interesting patterns that could facilitate the interpretion of the results from other analyses, they were thoroughly presented in this section. As a multivariate analysis procedure similar to MANOVA, Wilks' Lambda has been routinely reported (Tabachnick & Fidell's, 2001). The magnitude of Wilks' Lambda represents the proportion of variance of self- and other- ratings that is not accounted for by the demographic variables.

Table 6

Demographics and Self-Other Agreement with the U.S. Sample

	Self				Other			
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*
Self vs. Subordinate								
Gender	-.024	.031	.759	.448	.054	.036	1.493	.136
Race	.048	.045	1.064	.288	.003	.052	.056	.955
Age	.004†	.002	1.946	.052	-.002	.002	.732	.465
Education	.015	.033	.471	.638	.021	.038	.546	.585
R^2	.008				.004			
Wilks' Λ	.985							
Self vs. Peer								
Gender	-.009	.031	.293	.770	.042	.029	1.462	.144
Race	.052	.043	1.223	.222	-.034	.040	.839	.402
Age	.006**	.002	2.954	.003	.003†	.002	1.771	.077
Education	.011	.031	.342	.733	.027	.030	.914	.361
R^2	.013*				.008			
Wilks' Λ	.980*							

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

Demographic Effects in the U.S. Sample

Table 6 demonstrated demographics were not related to self- and subordinates-ratings, but were related to self- and peer-ratings jointly in the U.S. sample (Wilks' Λ = .980, $F(2, 763) = 1.947$, $p = .050$). A follow-up univariate analysis indicated that only self-ratings were affected ($R^2 = .013$, $F(4, 761) = 2.551$, $p = .038$). Particularly, older ratees tended to slightly overestimate themselves compared with subordinates. The overestimation effect weakened when their ratings were compared to peer-ratings, because peers also gave higher ratings to older ratees. Gender, race, and education were not related to individual ratings as well as self-other agreement in the U.S. sample.

As shown in Table 7, demographics also affected self-ratings and supervisor-ratings of leadership (Wilks' Λ = .974, $F(2, 815) = 2.708$, $p = .006$), and self-ratings and supervisor-ratings of performance (Wilks' Λ = .979, $F(2, 815) = 2.203$, $p = .025$). The four demographic variables accounted for a significant portion of variance in self-ratings ($R^2 = .012$, $F(4, 813) = 2.376$, $p = .051$), with older ratees giving themselves higher ratings than their younger counterparts, resulting an overestimation effect. These demographic factors were also related to supervisor-ratings of leadership ($R^2 = .016$, $F(4, 813) = 3.263$, $p = .011$), but not related to supervisor-ratings of performance ($R^2 = .009$, $F(4, 813) = 1.920$, p = *n.s.*). Ratees with more education received higher ratings from their supervisors than did their counterparts who had less education. Because self-ratings were not related to education level, ratees with higher education

Table 7

Demographics and Self-Supervisor Agreement with the U.S. Sample

	Self				Other			
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*
Self vs. Supervisor								
Gender	-.018	.030	.517	.537	.069†	.038	1.846	.065
Race	.051	.042	1.220	.223	.022	.052	.420	.674
Age	.005**	.002	2.805	.005	.003	.002	1.425	.155
Education	.009	.031	.296	.767	.109**	.039	2.831	.005
R^2	.012†				.016*			
Wilks' Λ	.974**							
Self vs. Performance								
Gender	-.018	.030	.517	.537	.083†	.044	1.871	.062
Race	.051	.042	1.220	.223	.014	.062	.220	.826
Age	.005**	.002	2.805	.005	-.001	.003	.190	.850
Education	.009	.031	.296	.767	.096*	.046	2.099	.036
R^2	.012†				.009			
Wilks' Λ	.979*							

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

appeared to be more accurate estimators than those with lower education. In other words, ratees with lower education tended to be over-estimators. Overall, the variance accounted for by demographic variables was small and would not affect MSF ratings substantially.

Demographic Effects in the Chinese Sample

Table 8 indicates that gender and education were significantly related to self- and subordinate-ratings jointly in the Chinese sample (Wilks' Λ = .949, $F(2, 346) = 3.013$, $p = .006$). Further univariate analyses revealed that the effects were resulted from self-ratings ($R^2 = .049$, $F(3, 345) = 5.870$, $p = .006$), not from subordinate-ratings ($R^2 = .006$, $F(3, 345) = .682$, $p = n.s.$). Male ratees rated themselves higher than their female counterparts while subordinate-ratings were not related to the ratee's gender, making males over-estimators. Ratees who had more education rated themselves higher than those who had less while subordinate-ratings were not related to the ratee's education level, making ratees with more education over-estimators. Self- and peer-ratings together were strongly influenced by demographic factors in the Chinese sample (Wilks' Λ = .933, $F(2, 392) = 4.559$, $p < .001$), due to self-ratings ($R^2 = .059$, $F(3, 391) = 8.219$, $p < .001$), and peer-ratings to a lesser extent ($R^2 = .017$, $F(3, 391) = 2.253$, $p = .082$). Similar to the self-subordinate comparison, males and ratees with more education tended to be over-estimators compared with peers. Older ratees gave themselves higher ratings than their younger counterparts, but they also received higher ratings from their peers. The overestimation effect was mitigated.

Table 8

Demographics and Self-Other Agreement with the Chinese Sample

	Self				Other			
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*
Self vs. Subordinate (*n* = 349)								
Gender	-.139**	.048	2.894	.004	-.012	.052	.239	.811
Age	.005	.004	1.453	.147	.004	.004	1.021	.308
Education	.094*	.047	1.986	.048	.043	.051	.841	.401
R^2	.049**				.006			
Wilks' Λ	.949**							
Self vs. Peer (*n* = 395)								
Gender	-.105*	.047	2.210	.028	.017	.039	.434	.665
Age	.012**	.004	3.203	.001	.007*	.003	2.127	.034
Education	.093*	.047	1.986	.048	.054	.039	1.382	.168
R^2	.059**				.017†			
Wilks' Λ	.933**							

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

The finding that ratees with more education were over-estimators compared with both subordinates and peers was inconsistent with what had been found in the

Table 9

Demographics and Self-Supervisor Agreement with the Chinese Sample

	Self				Other			
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*
Self vs. Supervisor								
Gender	-.099*	.045	2.189	.029	.104*	.047	2.239	.026
Age	.012**	.004	3.265	.001	.005	.004	1.310	.191
Education	.090*	.044	2.032	.043	.107*	.046	2.326	.021
R^2	.056**				.025*			
Wilks' Λ	.921**							
Self vs. Performance								
Gender	-.099*	.045	2.189	.029	.138*	.055	2.498	.013
Age	.012**	.004	3.265	.001	.001	.004	.237	.813
Education	.090*	.044	2.032	.043	.166**	.055	3.049	.002
R^2	.056**				.033**			
Wilks' Λ	.911**							

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed).

U.S. sample, where ratees with less education were over-estimators.

In the Chinese sample, demographic factors had a strong influence on self-ratings and supervisor-ratings of leadership (Wilks' Λ = .921, $F(2, 425)$ = 5.949, $p < .001$), and self-ratings and supervisor-ratings of performance (Wilks' Λ = .911, $F(2, 425) = 6.738$, $p < .001$). Table 9 indicates that both self- and supervisor-ratings were significantly related to the demographics. The three demographic variables accounted for 5.6% of the variance in self-ratings ($F(3, 424) = 8.324$, $p < .001$), 2.5% in supervisor-ratings of leadership ($F(3, 424) = 3.559$, $p = .014$), and 3.3% in supervisor-ratings of performance ($F(3, 424) = 4.860$, $p = .002$). Males gave themselves higher ratings than females but their supervisors rated females higher than males. As a result, males became severe over-estimators while females appeared to be accurate estimators. Older ratees also rated themselves higher than their younger counterparts, but supervisor-ratings were not related to age. Therefore, older ratees appeared to be over-estimators compared with supervisors. Ratees with more education gave themselves higher ratings than their counterparts and received higher ratings from their supervisors as well. The overestimation effect was not present.

Demographic Effects in the GDT Sample

Table 10 illustrates that demographic factors were significantly related to self- and subordinate-ratings jointly in the GDT sample, (Wilks' Λ = .982, $F(2, 795) = 2.378$, $p = .027$). The demographic effects were related to both self-ratings ($R^2 = .010$, $F(3, 794) = 2.759$, $p = .041$) and subordinate-ratings ($R^2 = .011$, $F(3, 794) = 2.822$, $p = .038$). While older ratees gave themselves higher ratings, they also

Table 10

Demographics and Self-Other Agreement with the GDT Sample

	Self				Other			
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*
Self vs. Subordinate								
Gender	.015	.035	.419	.675	.047	.037	1.288	.198
Age	.006**	.002	2.871	.004	.006**	.002	2.709	.007
Education	-.002	.030	.077	.939	.011	.032	.347	.738
R^2	.010**				.011*			
Wilks' Λ	.982*							
Self vs. Peer								
Gender	.019	.035	.537	.582	.063*	.031	2.062	.040
Age	.007**	.002	3.542	< .001	.003	.002	1.588	.113
Education	.006	.030	.198	.843	-.005	.026	.186	.852
R^2	.015*				.008†			
Wilks' Λ	.979*							

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

received higher ratings of the same magnitude from their subordinates. Thus, no overestimation effect was observed. Demographics were also related to self- and

peer-ratings jointly in the GDT sample (Wilks' Λ = .979, $F(2, 831) = 2.906$, $p = .008$), due to both self-ratings ($R^2 = .015$, $F(3, 830) = 4.210$, $p = .006$) and peer-ratings ($R^2 = .008$, $F(3, 830) = 2.094$, $p = .099$). Females gave themselves ratings close to those of males, but their peers rated them higher than males. As a result, females would be more accurate estimators than males, who instead appeared to be over-estimators. Older ratees gave themselves higher ratings than did their younger counterparts, while peer-ratings remained impartial. Thus, older ratees appeared to be over-estimators. Peers rated female ratees higher than males but self-ratings were not related to gender. Males seemed to be over-estimators compared with females. Neither self-ratings nor peer-ratings were related to education in the GDT sample.

Demographics affected self- and supervisor-ratings of leadership only marginally (Wilks' Λ = .986, $F(2, 868) = 2.090$, $p = .052$), but they affected self-ratings and supervisor-ratings of performance significantly (Wilks' Λ = .979, $F(2, 868) = 3.023$, $p = .006$). Table 11 indicates that self-ratings were significantly related to the demographics ($R^2 = .012$, $F(3, 867) = 3.586$, $p = .013$) while the relationship between demographic variables and supervisor-ratings of performance was only marginal ($R^2 = .008$, $F(3, 867) = 2.194$, $p = .087$). Older ratees tended to over-estimate themselves and the degree of overestimation became more pronounced as their supervisors gave them lower performance ratings than those younger ratees. Gender and education level were not related to both self- and supervisor-ratings.

Table 11

Demographics and Self-Supervisor Agreement with the GDT Sample

	Self				Other			
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*
Self vs. Supervisor								
Gender	.018	.034	.543	.587	.042	.037	1.132	.258
Age	.006**	.002	3.266	.001	.001	.002	.246	.805
Education	.006	.029	.215	.830	.025	.032	.808	.419
R^2	.012**				.002			
Wilks' Λ	.986							
Self vs. Performance								
Gender	.018	.034	.543	.587	.053	.046	1.160	.247
Age	.006**	.002	3.266	.001	-.005*	.003	2.039	.042
Education	.006	.029	.215	.830	-.026	.039	.676	.499
R^2	.012**				.008†			
Wilks' Λ	.979*							

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

Overall, the demographic effects were present in all three samples. Their effects were the most pronounced in the Chinese sample, with all three variables

being related to self- and other-ratings. The GDT sample was the least affected by demographic factors, as gender and education were only related to self- and other-ratings in few cases. Among the three demographic variables (gender, age, and educational level), age had the most consistent effects on self-ratings, represented by an overestimation effect in older ratees. Education level affected supervisor-ratings in the U.S. and the Chinese samples but not in the GDT sample. Gender affected both self- and supervisor-ratings in the Chinese sample. Therefore, demographic variables seemed to affect ratings disparately in different cultural settings and across different perspectives. As a result, demographic variables were included in following polynomial analyses.

Polynomial Regression

Polynomial regression analyses were conducted in three phases. First, two separate polynomial regressions were run for self-subordinate and self-peer ratings, with all participants included. Second, two similar polynomial regressions were run, but with sample source as moderator. Significant cross-product terms that included the dummy variables would suggest the presence of moderating effects. Third, polynomial regressions were conducted independently for separate samples.

Omnibus Relationships

With all participants included, separate polynomial regressions were run for self-subordinate and self-peer ratings, respectively (see Table 12). For self-

Table 12

Polynomial Regressions with All Participants

	Self-Subordinate (*n* = 1,862)				Self-Peer(*n* = 1,995)			
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Constant	4.032	.018	3.970	.025	4.040	.017	3.954	.024
Self	.048	.032	.057†	.032	.052†	.030	.062*	.030
Sub	.344**	.030	.334**	.031	.502**	.033	.491**	.033
Self squared	-.032	.056	-.017	.056	-.016	.051	-.002	.050
Self * Other	.070	.072	.065	.072	.038	.083	.030	.083
Other squared	-.037	.049	-.040	.049	-.129*	.063	-.115†	.063
Gender			.073**	.019			.090**	.018
Age			-.002	.002			-.004*	.002
Education			.032	.027			.055*	.025
R^2	.080**		.088**		.119**		.133**	

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

subordinate agreement, the polynomial regression equation was significant, with $R^2 = .080$, $F(5, 1856) = 32.148$, $p < .001$. The polynomial regression equation for self-peer agreement was also significant, with $R^2 = .119$, $F(5, 1989) = 53.579$, $p < .001$. Demographic variables accounted for some additional variance but had

little impact on the size of parameters in the regression equations. For self-subordinate agreement, self-ratings predicted performance marginally well, with B = .057, t(1853) = 1.753, p = .080. In contrast, subordinate-ratings displayed a strong relationship with performance, with B = .334, t(1853) = 10.935, p < .001. However, the quadratic terms were not significant. The response surface displayed in Figure 4a was generally flat. Along the line of perfect agreement (X = Y), performance increased linearly as self- and other-ratings went from the low end to the high end, with a_1 = .391, F(1, 1856) = 102.355, p < .001, and a_2 = .008, F(1, 1856) = .000, p = *n.s.*. The line of complete disagreement (X = -Y) had a negative slope, with a_3 = -.277, F(1, 1856) =35.540, p < .001, and appeared slightly convex, with a_4 = -.122, F(1, 1856) = 1.302, p = *n.s.*, suggesting that over-estimators tended to have lower performance than under-estimators.

For self-peer agreement, performance was related to both self-ratings (B = .062, t(1986) = 2.984, p = .037) and peer-ratings (B = .491, t(1986) = 14.819, p < .001). The quadratic term of peer ratings was also related to performance, but only marginally (B = -.115, t(1986) = 1.827, p = .068). Other quadratic terms were not significant. The response surface in Figure 4b showed a slightly convex shape. Along the line of perfect agreement, in agreement at higher levels was related to higher performance than in agreement at lower ratings of behavior, with a_1 = .553, F(1, 1989) = 197.837, p < .001. The relationship appeared linear, with a_2 = -.087, F(1, 1989) = 1.282, p = *n.s.*. Over-estimators had a poorer performance than under-estimators, as demonstrated by the negative slope of

Figure 4. Relationships between self-other agreement and performance.

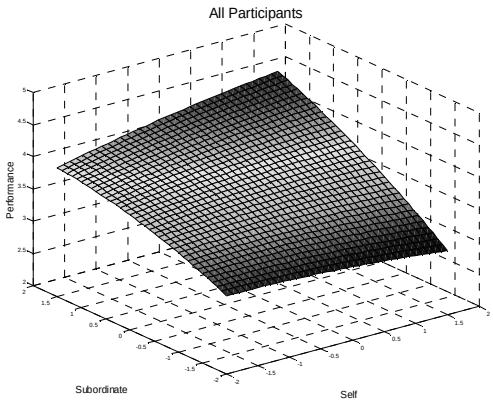


Figure 4a

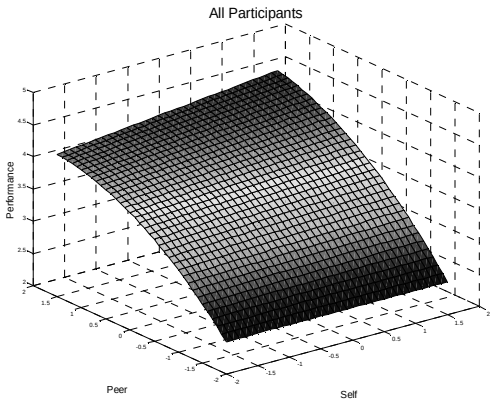


Figure 4b

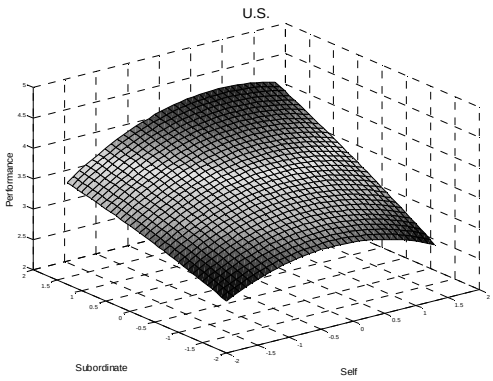


Figure 4c

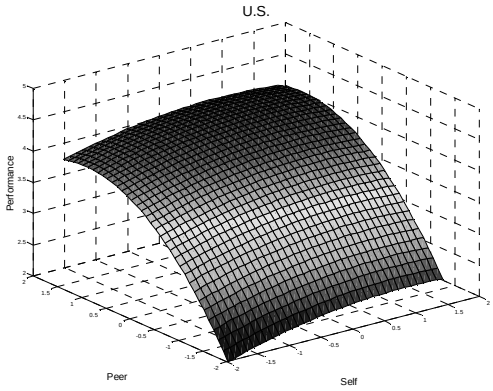


Figure 4d

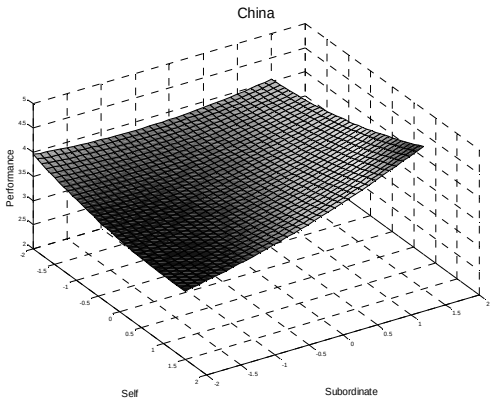


Figure 4e

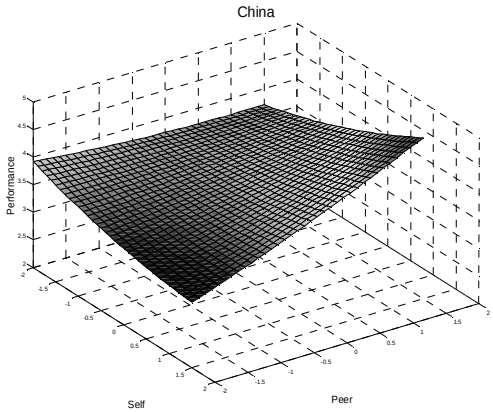


Figure 4f

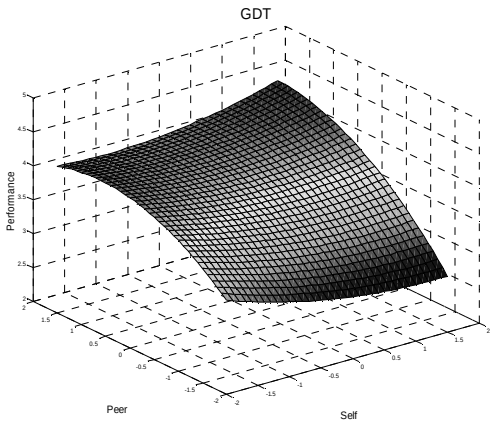


Figure 4g

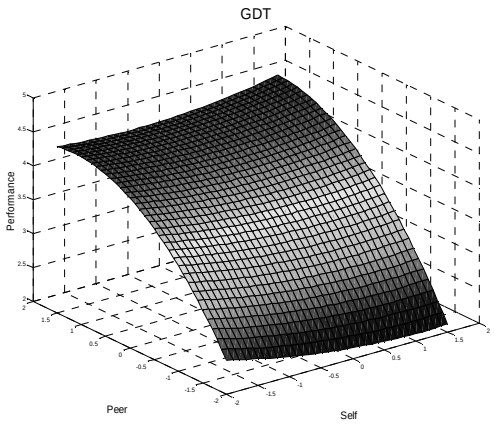


Figure 4h

the line of complete disagreement, with a_3 = -.429, F(1, 1989) = 197.837, p < .001. The line appeared slightly convex, with a_4 = -.147, F(1, 1989) = 1.801, p = *n.s.*.

Moderated Polynomial Regressions

To investigate whether the pattern of relationships between self-other agreement and performance was different across samples, moderated polynomial regressions were conducted, with sample source as moderator. Because very few terms in the regression equations turned out significant, tables are not presented here. The results of the first moderated polynomial regression showed that the pattern of relationship between self-subordinate agreement and performance was similar in all three samples, since adding the interaction terms with sample source contributed only .004 to R^2 with F(10, 1844) = .804, p = *n.s.*. The results of the second moderated polynomial regression showed some significant differences in the pattern of relationships between self-peer agreement and performance, with R^2 = .008, F(10, 1977) = 1.835, p = .050. The

slope for peer-ratings in the Chinese sample was .224 points lower than that in the U.S. sample ($t(1977) = 2.234$, $p = .026$), indicating that peer-ratings in the Chinese sample were not as a strong predictor for performance as those in the U.S. sample. Since a significant moderating effect was present in the second moderated polynomial regression, it is worthwhile to conduct independent polynomial regressions on separate samples.

Polynomial Regression Analysis by Sample

Two polynomial regressions were conducted for each sample. The race variable was not included because it was not related to self-other agreement in the U.S. sample and was not included in the Chinese and the GDT samples either. Removing the variable from polynomial regressions was a reasonable act. To facilitate comparisons between samples, separate response surfaces were drawn (see Figure 4).

Polynomial regression results for the U.S. sample. For the U.S. sample, both polynomial regression equations were significant (see Table 13). Self- and subordinate-ratings accounted for 8.8% of the variance in performance ($F(5, 709) = 13.628$, $p < .001$), while self- and peer-ratings accounted for 12.1% of it ($F(5, 760) = 20.827$, $p < .001$). Subordinate-ratings were significantly related to performance ($B = .362$, $t(706) = 7.008$, $p < .001$), but neither the term of self-ratings nor the quadratic terms were significant. Nevertheless, the response surface (see Figure 4c) indicated that self-ratings were positively related to performance when subordinate-ratings were high. Performance increased

Table 13

Polynomial Regressions with the U.S. Sample

	Self-Subordinate (n = 715)				Self-Peer(n = 766)			
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Constant	4.121	.031	4.040	.072	4.150	.029	4.026	.067
Self	.041	.057	.042	.057	.081	.052	.086†	.052
Sub	.365**	.052	.362**	.052	.488**	.056	.482**	.056
Self squared	-.157	.102	-.144	.102	-.057	.092	-.053	.092
Self * Other	.067	.133	.075	.133	-.012	.151	.012	.150
Other squared	-.022	.081	-.024	.081	-.243*	.101	-.235*	.101
Gender			.044	.046			.071†	.043
Age			.000	.003			-.003	.003
Education			.067	.048			.086*	.044
R squared	.088**		.091**		.121**		.129**	

Notes: ** p < 0.01 (2-tailed), * p < .05 (2-tailed), † p < .10 (2-tailed).

linearly along the line of perfect agreement as self-subordinate ratings increased, with a_1 = .404, F(1, 709) = 38.244, p < .001, and a_2 = -.093, F(1, 709) = .678, p = *n.s.*. Performance of under-estimators was higher than that of over-estimators along the line of complete disagreement, with a_3 = -.320, F(1, 709) = 14.145, p

< .001. The line appeared slightly convex, with a_4 = -.243, F(1, 709) = 1.143, p = *n.s.*. The slightly convex line suggested that performance dropped at an increasing rate when the degree of overestimation increased, while extreme under-estimators seemed to have poor performance. Peer-ratings were also significantly related to performance (B = .482, t(757) = 8.584, p < .001), but self-ratings were only marginally related (B = .086, t(757) = 1.668, p = .096). The quadratic term for peer-ratings were also significant (B = -.235, t(757) = 2.348, p = .019), suggesting a convex-shaped response surface. Examination of surface features (see Figure 4d) suggested that performance decreased as self-ratings and peer-ratings went from high to low and decreased more rapidly when the ratings become lower along the line of perfect agreement, with a_1 = .553, F(1, 760) = 70.830, p < .001 and a_2 = -.276, F(1, 760) = 3.452, p = .064. Along the line of complete disagreement, under-estimators achieved higher performance than over-estimators and performance became successively lower as lack of agreement increased, with a_3 = -.396, F(1, 760) = 23.285, p < .001 and a_4 = -.300, F(1, 760) = 1.440, p = *n.s.*. The pattern of relationships is similar to that found in Atwater et al. (2005) with the U.S. sample.

Polynomial regression results for the Chinese sample. For the Chinese sample, both polynomial regression equations were significant (see Table 14). Self- and subordinate-ratings accounted for 8.5% of the variance in performance (F(5, 343) = 6.342, p < .001), but self- and peer-ratings accounted for only 6.3% of the variance (F(5, 389) = 5.241, p < .001), which was substantially lower than

Table 14

Polynomial Regressions with the Chinese Sample

	Self-Subordinate (*n* = 349)				Self-Peer(*n* = 395)			
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Constant	3.840	.040	3.742	.049	3.834	.038	3.724	.047
Self	.109	.069	.125†	.070	.128*	.063	.146*	.063
Sub	.291**	.065	.285**	.065	.268**	.077	.254**	.076
Self squared	.055	.122	.063	.121	.039	.106	.031	.104
Self * Other	.118	.160	.096	.159	.160	.189	.136	.186
Other squared	.057	.108	.068	.107	.021	.149	.065	.147
Gender			.152**	.060			.172**	.057
Age			.000	.005			-.002	.005
Education			.132**	.058			.144**	.056
R^2	.085**		.115**		.063**		.099**	

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

that in the U.S. sample. Subordinate-ratings were significantly related to performance ($B = .285$, $t(340) = 4.418$, $p < .001$), while self-ratings were marginally related ($B = .125$, $t(340) = 1.797$, $p = .073$). Other terms were not significant. The response surface (see Figure 4e) appeared rather flat.

Examination of surface features suggested that a_1 was .410 (F(1, 343) = 25.287, p < .001) and a_2 was .227 (F(1, 343) = 1.754, p = *n.s.*), representing a positively sloped but slightly concave line along the line of perfect agreement. Thus, performance decreased as self-subordinate ratings moved downwards from the high end but leveled off at the low end. Along the line of complete disagreement, under-estimators achieved higher performance than accurate estimators and over-estimators, with a_3 = -.160, F(1, 343) = 2.852, p = .092) and a_4 = .037, F(1, 343) = .001, p = *n.s.*.

Performance was significantly related to both self-ratings (B = .146, t(386) = 2.298, p = .022) and peer-ratings (B = .254, t(384) = 3.330, p = .001), but other terms were not significant. Examination of surface features (also see Figure 4f) suggested that a_1 was .400 (F(1, 389) = 22.019, p < .001) and a_2 was .232 (F(1, 389) = 1.368, p = *n.s.*). Therefore, the line of perfect agreement was positively sloped and slightly concave. Similar to self-subordinate agreement, Figure 4f illustrated that the line of perfect agreement also leveled off at the lower end. The line of complete disagreement had a modest slope and was virtually linear, with a_3 = -.106, F(1, 389) =1.549, p = *n.s.*, and a_4 = -.040, F(1, 389) = .092, p = *n.s.*, suggesting that performance of under-estimators were not different from that of over-estimators.

The response surfaces corresponding to these two polynomial regressions in the Chinese sample were quite similar to each other, with the line of perfect agreement peaking at the higher end and leveling off at the lower end. Because

the line of complete disagreement had a modest slope and was linear, overestimation only corresponds to slightly poorer performance, leading to the rejection of Hypothesis 3. Since under-estimators in the Chinese sample tended to have higher performance than those in the U.S. sample, Hypothesis 4 was supported.

Polynomial regression results for the GDT sample. Polynomial regression results for the GDT sample are present in Table 15. Self- and subordinate-ratings accounted for 6.8% of the variance in performance ($F(5, 792) = 11.484$, $p < .001$), while self- and peer-ratings accounted for 13.5% of the variance ($F(5, 828) = 25.874$, $p < .001$). Subordinate-ratings were significantly related to performance ($B = .315$, $t(789) = 6.786$, $p < .001$), but self-ratings were not. The quadratic term of subordinate ratings was marginally related to performance ($B = -.131$, $t(789) = 1.722$, $p = .086$), suggesting a convex shape (see Figure 4g). Since self-ratings were not related to performance, self-other agreement would be less relevant. The parameters related to the response surfaces were not reported. Figure 4g indicates that accurate estimators with high self-ratings tended to be high performers and those who overestimated themselves the most had the lowest performance.

Peer-ratings were significantly related to performance ($B = .545$, $t(825) = 10.980$, $p < .001$), but self-ratings were not. The quadratic term of peer-ratings was marginally related to performance ($B = -.181$, $t(825) = 1.838$, $p = .066$), suggesting a convex shape (see Figure 4h). The parameters related to response

Table 15

Polynomial Regressions with the GDT Sample

	Self-Subordinate (*n* = 798)				Self-Peer(*n* = 834)			
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Constant	4.038	.027	4.057	.035	4.038	.026	4.039	.033
Self	.011	.048	.021	.048	-.007	.044	.006	.044
Sub	.307**	.046	.315**	.046	.543**	.050	.545**	.050
Self squared	.050	.080	.043	.080	.040	.073	.034	.074
Self * Other	.074	.103	.059	.103	.043	.119	.026	.119
Other squared	-.127†	.076	-.131†	.076	-.197*	.098	-.181†	.098
Gender			.033	.046			.020	.043
Age			-.007**	.003			-.006**	.002
Education			-.056	.039			-.017	.037
R^2	.068**		.079**		.135**		.143**	

Notes: ** $p < 0.01$ (2-tailed), * $p < .05$ (2-tailed), † $p < .10$ (2-tailed).

surfaces were not reported, due to the same aforementioned reason. Figure 4h indicated that performance was solely determined by peer-ratings and the relationship between peer-ratings and performance was strong.

Figures 4g and 4h shared many commonalities. Both subordinate- and peer-ratings were related to performance and took a convex shape. However, because self-ratings were not related to performance, there was little evidence showing underestimation and overestimation would result in poorer performance in the GDT sample than in the U.S. sample. Therefore, both Hypotheses 7 and 8 were not supported.

Summary

The pattern of means and standard deviations was similar across the three samples. However, the magnitude of the means in the Chinese sample was consistently lower than that in the other two samples. The correlations between other-ratings were stable across samples but those between self- and other-ratings were less stable. The self-other correlations were the highest in the Chinese sample but were the lowest in the GDT sample. While the SEQ results did identify some correlations of different sizes between self- and other-ratings between the Chinese and the GDT samples, no differences were present between the U.S. and the Chinese samples, and between the U.S. and the GDT samples, leading to the rejection of Hypotheses 1 and 5. The correlations between other-ratings in the Chinese sample were not different from those in the U.S. sample, except that self-ratings correlated significantly lower with supervisor-ratings of leadership in the Chinese sample, lending partial support to Hypothesis 2. Since the correlations between other-ratings in the GDT sample were not different from those in the U.S. sample, Hypothesis 6 was not supported.

Multivariate regression analyses revealed demographic variables were related to self-other agreement. Among all four perspectives, self-ratings were the most susceptible to demographic influences. Among all demographic variables, age demonstrated the most consistent effects on self-ratings across all three samples. Among the samples, the Chinese sample was affected by demographics the most while the GDT sample the least. Thus, demographic effects on self-other agreement varied across perspectives and samples. As demographic effects were present, demographic variables were included in subsequent polynomial regression equations.

The polynomial regression results indicated disparate relationships between self-other agreement and performance across the three samples. The results for the U.S. sample generally replicated those found in other previous studies with different U.S. samples. The response surfaces in the Chinese sample were flatter than those in the U.S. sample and the overestimation effects appeared only in some extreme situations, so Hypothesis 3 was not supported. In contrast, under-estimators in the Chinese culture tended to have slightly higher performance than accurate estimators, lending support to Hypothesis 4. For the GDT sample, neither overestimation nor underestimation was related to performance, so Hypotheses 7 and 8 were not supported.

CHAPTER FIVE

DISCUSSION

Means and Standard Deviations

The descriptive statistics indicated that the pattern of means of different perspectives was similar across all three samples, with supervisor-ratings of performance being the highest, followed by self- and subordinate-ratings, and peer-ratings being one of the lowest. Supervisors could differentiate ratees the best, indicated by the largest standard deviations they had. Peers had the smallest standard deviations, making them less discriminating raters. The observation that peers had the lowest means and standard deviations is consistent with what has been found in the Atwater et al. (2005) and Ostroff et al. (2004) studies, which also used the Profilor measure.

Ratings in the Chinese sample averaged consistently lower than those in the other two samples, particularly for peer- and supervisor-ratings. While it may imply a modesty tendency in Chinese, it may also reflect some true discrepancies in leadership competencies between Chinese and American managers. Chinese mangers are generally younger, more poorly trained, and less familiar with empowering, facilitating, and team-based leadership which has been practiced widely in the West. Management in China is still underdeveloped and team-related research and practice have lagged far behind the U.S.

In the Chinese sample, the mean of self-ratings was similar to that of subordinate-ratings, but substantially higher than those of peer- and supervisor-ratings. Therefore, a leniency bias suggested by Fahr, Dobbins and Cheung (1991), which states Chinese employees tend to rate themselves less favorably than do their supervisors, is not present in this study. The pattern observed is consistent with that in Furnham and Stringfield (1994) and Goodstein, Stringfield and Zevulun (1990). Significant changes have taken place in China during the past 15 years, which may contribute to the shift in rating style. In addition, ratees working in multinational companies may have changed their behavior to accommodate westernized organizational structure and culture.

The standard deviation of self-ratings in the Chinese sample was larger than that in the U.S. sample but close to those of subordinate-ratings and supervisor-ratings of leadership. Chinese ratees are comfortable to give themselves both high and low ratings. As MSF is mainly used for developmental purposes, rating oneself low may reflect a drive to learn. People from a collectivist culture may pay more attention to self-improvement in order to meet role obligations (Heine et al., 2001). The standard deviations of supervisor-ratings in the Chinese sample were found to be smaller than those in the U.S. sample, which lends some support to Atwater et al.'s (2005) speculation that people from the collectivist culture may give less discriminating ratings to others.

Correlations between Ratings

A general pattern of correlations emerges across samples. First, correlations between other-ratings are generally larger than those between self- and other-ratings. Second, correlations between other-ratings seem very stable across samples, within a narrow range of high .20s and low .40s. Third, correlations between self- and other-ratings vary considerably across samples. The self-other correlations in the Chinese sample were consistently higher than those in the U.S. sample, which in turn were higher than those in the GDT sample. It seems that culture's influence on consensus between other-ratings is trivial in comparison to self-other agreement.

Although the correlations between self- and other-ratings were consistently higher in the Chinese sample than those in the U.S. sample, the follow-up EQS analysis suggested the differences were not significant. Therefore, Hypothesis 1 is not supported. However, the differences between the Chinese and the GDT samples were found significant, which implies that the differences between the Chinese and the U.S. samples could be something more than errors. A possible reason for not finding the differences significant may lie in the population the Chinese sample represented. Most participants in the Chinese sample worked for multinational companies. If participants from local Chinese companies are enlisted, the differences may actually turn out significant.

Hypothesis 2 speculated that the correlations between other-ratings would be lower in the Chinese sample than that in the U.S. sample. The hypothesis is

only partially supported, with the correlations between peer-ratings and supervisor-ratings of leadership in these two samples being significantly different. The lower correlation between peer-ratings and supervisor-ratings of leadership in the Chinese culture is not caused by in-group practices, since the mean of peer-ratings was as low as that of supervisor-ratings. It is not caused by the high power distance as well, since subordinates agreed with peers and supervisors reasonably well. The lower correlation is not caused by the restriction of range either, though the standard deviations of peer- and supervisor-ratings were the lowest within each perspective. The GDT sample also displayed smaller standard deviations than the U.S. sample in both ratings but the correlation was slightly higher than that in the U.S. sample. The reason for the lower correlation between peer- and supervisor-ratings in the Chinese sample may lie in a lack of team practices at the higher level of the organization in China. Chinese supervisors may be less likely to act as team leaders. As a result, supervisors and peers may observe different behaviors of the same ratee and come up with different ratings.

With the U.S. participants, peer-rating has established itself as a more reliable and valid measure than self-rating and subordinate-rating (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988), and self-peer agreement has a stronger relationship with performance than self-subordinate agreement (e.g., Atwater et al., 2005; Atwater et al., 1998; Ostroff et al., 2004). However, in the Chinese sample, peer-ratings correlated less with supervisor-ratings and the relationship between self-peer agreement and performance was less pronounced

than that with self-subordinate agreement. Self- and peer-ratings usually account for more variance in performance than do self- and subordinate-ratings, consistent with the general finding that the correlation between peer-ratings and supervisor-ratings is the highest among all correlation coefficients. However, the opposite was found in the Chinese sample. The subordinate perspective seems to be more important than the peer perspective in the Chinese setting.

Therefore, the correlations between ratings in the Chinese sample are similar to those in the U.S. sample. There are several reasons for not finding the expected relationships in the Chinese sample. First, the majority of the Chinese participants are from multinational companies and people may have been socialized to behave differently under the influence of the well established organizational structure and culture. Second, Chinese culture is under change and the new generation of mangers in China is becoming more individualist than older managers (Ralston et al., 1999). For managers working in multinational companies, this trend can be even more prominent. Third, other processes may exist and offset the effects expected. For example, collectivists may have fewer self-linked memories than individualists and generally know more about others than about self (Triandis, 1994). Fourth, from the measurement perspective, aggregating ratings may have cancelled out the effects of in-group practices. Because self-, subordinate-, and peer-ratings are all related to performance, and also because in-group practices and power distance seem to have little impact

on ratings, using MSF in the Chinese culture, at least for multinational companies operating in China, seems appropriate.

Although the self-other correlations in the GDT sample appeared consistently lower than those in the U.S. sample, the EQS results suggested the differences were not significant, leading to the rejection of Hypothesis 5. However, the differences between the Chinese and the GDT samples indicated that the latter may indeed had lower self-other correlations. A more careful selection of participants for the GDT sample may be able to help achieve the results expected. The correlations among other-ratings in the GDT sample were not different from those in the U.S. sample either, lending support to Hypothesis 6. Therefore, raters can reach consensus about a ratee's performance satisfactorily despite cultural, language, and geographical barriers, suggesting that using MSF in GDTs is quite plausible. The structured work environment characterized by GDTs may have contributed to overcoming the barriers.

Person perception theories (e.g., Kenny, 1991, 1994) provide some important insights how accuracy is possible in such adversarial situations. When a rater works with a ratee for some time, he/she will observe a large amount of the ratee's behaviors and give ratings based on the behaviors. Consensus will be influenced by the amount of behavior observed, the degree of overlap of behavior between raters, behavioral consistency of the ratee, raters' meaning systems, communication between raters, and other unique impressions formed through interactions. While consensus may be jeopardized by fewer behaviors

being observed, dissimilar meaning systems, and constrained communication, a more structured working environment (using conference calls, emails) typical in GDTs may help to improve consensus. As shown in assessment centers, the more structured the situation, the more accurate the rating. It also suggests that observing relevant behaviors may be more important than observing more behaviors (Landy & Farr, 1980).

Demographics Effects

Among the three demographic variables, age had the most consistent effects on self-ratings, with older ratees tending to give themselves higher ratings than their younger counterparts. Since other-ratings were generally not affected by age, older ratees would be more likely to overestimate themselves. This finding is consistent with what has been reported in Ostroff et al. (2004). The overestimation effect may reflect that older managers are less receptive to feedback in these cultures (Ryan, Brutus, Gregurus, & Hakel, 2000), and that they rely more on self-concept for ratings. Interestingly, age was positively related to supervisor-ratings of leadership but negatively to supervisor-ratings of performance in the GDT sample. It seems that older managers have the same level of leadership as their younger counterparts, but their performance is not as high as the latter. Older managers may have disadvantages to handle heavy workload, time zone differences, and extended travel that are common in GDTs.

Strong demographic effects on self-other agreement were present in the Chinese sample. Demographic variables accounted for 5.6% of variance in self-

ratings, which was disproportionately higher that in other-ratings, suggesting self-ratings in China are severely confounded with various stereotypes. Demographic variables also played an important role in supervisor-ratings, with more than 2.5% of variance being accounted for. In contrast, subordinate- and peer-ratings were less affected. The strong demographic effects present in self- and supervisor-ratings suggest that the collective self-concept has been called up in the ratings. It has been suggested Chinese pay more attention context than content (e.g., Triandis, 2000), so they may rely less on self-concept for ratings than their individualist counterparts. This speculation seems to be unfounded. Collectivists are equally likely to call up self-concept when ambiguities arise, except that they tend to leverage the collective one.

In the Chinese sample, males gave themselves higher ratings than females, while subordinate- and peer-ratings were not related to gender. Therefore, males were over-estimators compared with subordinates and peers while females were accurate estimators, consistent with Atwater and Yammarino's (1997) and Ostroff et al.'s (2004) propositions. Females have accurate self-ratings because they are more responsive to feedback than males, who usually are more assertive, confident, and ignorant of feedback (Roberts & Nolen-Hoeksema, 1994). However, in the Ostroff et al. study, the overestimation effect in males was resulted from lower other-ratings, not from different self-ratings. It seems that under the same overestimation effect there exist different dynamisms. Interestingly, both dynamisms were present when supervisor-ratings

were considered in the Chinese sample. While males rated themselves higher than females, their supervisor-ratings gave them lower ratings than females. It is unclear why supervisors in the Chinese sample gave higher ratings to females than males. One possible reason is females have better foreign language skills than males, which makes them look more competent when working in multinational companies. Future studies should investigate whether the gender effects on supervisor-ratings reflect real performance differences between males and females or just are artifacts.

Education level was also related to self-other agreement in the Chinese sample. Managers with more education rated themselves higher than those with less education. However, their subordinates and peers did not rate them differently. Managers with more education were over-estimators compared with subordinates and peers. In contrast, supervisor-ratings were positively related to education level in the U.S. sample, but self-ratings were not. Managers with less education seemed to be over-estimators in the U.S. sample, supporting the notion that mangers with more education have better analytical skills and thus are able to achieve greater self-other agreement (Fletcher & Baldry, 2000). Therefore, different dynamisms may exist in the Chinese and the U.S. samples regarding the role of education level in self-other agreement. The tendency to assign ratees with more education higher ratings may originate from an entrenched and long-lasting belief in education in the Chinese culture.

Therefore, demographic effects in the Chinese sample appear more salient that those in the U.S. sample, which may indicate that the collective self-concept dominant in the Chinese culture has a strong influence on ratings, especially for self-ratings and supervisor-ratings. According to the framework proposed in this study, self-concept will be called up to construe ratings when a behavior has been observed repeatedly, when a behavior cannot be recalled readily, or when inconsistent behaviors are observed. For Chinese supervisors, there are three possible reasons for showing strong demographic effects. First, they may observe fewer behaviors of the ratee due to high power distance. Second, they are more likely to receive conflicting information about a behavior conducted by the ratee from others due to in-group practices. Finally, they usually have limited time to fill out the MSF questionnaire, in fear of increasing the cost of being indecisive (Fiske, 1993). For Chinese ratees, often times they must act differently to meet sometimes incommensurable expectations and thus their behaviors become less consistent. As a result, they may experience more ambiguities than their U.S. counterparts in the process of rating own behaviors. In addition, it has been suggested that collectivists have fewer self-linked memories than individualists (Triandis, 1994) and their self-concepts are less consistent (Markus & Kitayama, 1991). When ambiguities arise, people from the collectivist culture will either base their self-ratings on behaviors or resort to their collective self-concepts. Therefore, the salient demographic effects present in the Chinese culture may suggest that the collective self-concept has been called up.

In contrast, ratings in GDTs seemed to be more objective than those in China. Self-ratings were not biased by gender and education, although older managers still tended to overestimate themselves. The degree of overestimation increased compared with supervisor-ratings, as supervisors gave younger managers higher ratings. With this exception, supervisor-ratings demonstrated few stereotypical biases. Therefore, it can be argued that ratings in GDTs are based more on behaviors observed than the Chinese case, thus supporting the proposition that GDTs provide a structured environment that facilitates behavior observation. However, some may argue for an alternative explanation. That is, demographic effects are less salient in GDTs because people from different cultures may have different stereotypes about gender and educational level.

Overall, demographic variables affect ratings extensively in the Chinese sample. High self-ratings are associated with older, more educated, and male ratees, and high supervisor-ratings are related to more education and females. In contrast, few demographic effects are present in GDTs. Ostroff, et al. (2004) found demographic effects were similar across self-subordinate, self-peer, and self-supervisor comparisons. However, this study shows the opposite. Substantial differences exist across both comparisons and samples. Moreover, the same overestimation effect may have different underlying dynamisms in different cultures. Therefore, research on demographics and self-other agreement will be more valuable when cultural influences are taken into account.

Self-Other Agreement and Performance

Self-other agreement was found to be related to supervisor-ratings of performance with all three samples combined. However, the relationships between self-other agreement and performance in this study appear weaker than those found in the Atwater et al. (2005) study. In this study, self-subordinate ratings accounted for about 8% of variance in performance while self-peer ratings accounted for about 12% of it. In the Atwater et al. study, the numbers were 23% and 32%, respectively. However, the numbers in this study are higher than those listed in the Ostroff et al. (2004) study, which are 5% and 11%, respectively. The differences may result from the selection of performance criterion. Ostroff et al. used the same performance criterion as this study, while Atwater et al. used a one-item scale that measured overall competence. As reported by Atwater et al., the one-item scale correlated .58 with the five-item scale used in this study. These two criteria are different in terms of the construct being measured, with the former measuring competency and the latter measuring results. The overall competency measure is more proximal to ratings of leadership and narrower than the performance measure. Using a broad measure such as overall performance may result in underestimation of a relationship due to bandwidth differences (Hogan & Roberts, 1996).

The relationship between self-other agreement suggested by MSF researchers (e.g., Atwater & Yammarino, 1997; Yammarino and Atwater, 1993) can be illustrated through self-peer agreement in the U.S. sample. Performance

decreases as self-peer ratings move from the high end to the low end and performance drops at an increasing rate as ratings become lower along the line of perfect agreement. Over-estimators receive lower performance ratings than under-estimators and performance decreases at an increasing rate as lack of agreement increases along the line of complete disagreement. The best performers will be accurate estimators with high self-ratings and the worst performers will be accurate estimators with low self-ratings. Extreme under-estimators have lower performance than accurate estimators. However, such an ideal pattern of relationship between self-other agreement and performance seems to only present in the U.S. sample.

Self-Agreement and Performance in China

In comparison to the U.S. and the GDT sample, the response surfaces in the Chinese sample were rather flat, which may suggest that Chinese raters are less discriminating in rating others (Atwater et al., 2005). Overestimation was only weakly related to lower performance ratings. Under-estimators didn't perform worse than accurate estimators and over-estimators. While accurate estimators with high self-ratings remained to be the best performers, extreme over-estimators who had very high self-ratings but very low other-ratings tended to have the poorest performance.

In the Chinese culture, underestimation can be associated with such Confucian virtues as humility and empathy in China. These virtues help to reduce interpersonal or inter-clique conflicts and maintain harmony in a group.

Maintaining a good interpersonal relationship with others is critical to achieve good performance in the Chinese culture. On the contrary, overestimation reflects an inflated self-concept, less willingness to accept negative feedback (Yammarino & Atwater, 1997), as well as poorer social skills (Colvin, Block & Funder, 1995), and could result in worse performance (Kwan et al., 2004). In the Chinese setting, having poorer social skills or focusing overly on himself/herself may prevent a manger from fulfilling role demands and meeting other's expectations, making it difficult to solicit others' support. As a result, performance may suffer considerably.

While underestimation corresponds to better performance in China, the overestimation effect may take place only in extreme conditions. As shown in Figures 4e and 4f, when subordinate- or peer-ratings were at their lower end, the relationship between self-ratings and performance became negative, suggesting a possible overestimation effect that overrode the self-rating effect. That is, the more self-ratings depart subordinate- or peer-ratings, the lower the performance, with extreme over-estimators (very high self-ratings but very low subordinate- or peer-ratings) having the lowest performance. In this sense, the overestimation effect in China is more pronounced than that in the U.S.. However, cautions should be taken as the polynomial regression equations are based on only 428 cases. The conclusion that extreme over-estimators will have the lowest performance in China is only tentative.

A weak leveling effect at the lower end of the line of perfect agreement was found for both self-subordinate and self-peer agreement in the Chinese sample. Accurate estimators who rate themselves low do not necessarily have low performance, which is quite unusual when compared to other samples in this study and the samples from individualist cultures (Atwater et al., 2005). Atwater et al. reported a leveling effect at the upper end of the line of perfect agreement for countries with low masculinity/individualist cultures such as France and Denmark. They speculated that people in the low masculinity culture are inclined to avoid competition and thus give high performers less discriminating ratings. They also speculated that collectivists will give less discriminating ratings. While this study shows that overall other-ratings in China are less discriminating than those in the U.S., they may be particularly poorer in differentiating low performers.

Self-Other Agreement and Performance in GDTs

In GDTs, self-ratings, the quadratic term of self-ratings, and the product of self-ratings and other-ratings were not significant. A weak relationship was found between self-ratings and performance in GDTs, which suggests that overestimation and underestimation would be less relevant. The weak relationship between self-ratings and performance resembles those found in European countries (Atwater et al., 2005). A follow-up analysis was conducted by breaking GDTs down into different sub-types. When a ratee was in one country while his/her subordinates or peers were in other countries, the relationship between self-ratings and performance remained significant. However, when a

ratee was with part of his/her team in one country and the remaining team in other countries, self-ratings were barely related to performance. If the findings with the sub-types can be replicated in future studies, a conclusion can be drawn that it is not the cultural differences between the rater and the ratee that cause inaccurate self-ratings. Instead, it is the existence of multiple meaning systems in a team that makes accurate self-ratings difficult to achieve.

Surrounded by multiple meaning systems, ratees face more challenges to collect feedback and make sense how they are performing, while other-raters are less affected. For ratees in GDTs, cultural, language and communication barriers make accurate feedback collection less convenient. Moreover, ratees in GDTs may experience more surprises than those in other settings because their behaviors are more likely to receive different reactions from other and result in different consequences. They may have more difficulty in interpreting the feedback cues and make sense of them. As a result, self-ratings in GDTs become less accurate than those in other settings. For other-raters, although their ratings are also affected by different meaning systems, the challenges are of less significance. Other-raters just need to focus on the ratee's behavior they observe, not necessarily all others' perceptions. Actually, the more structured work environment in GDTs may help to facilitate behavioral observation. Therefore, MSF can be applied to GDTs, based on the robust correlations found between other-ratings, but self-other agreement will be of less avail, due to its low relationship with performance.

Generalizability of the Self-Other Agreement Effect

MSF researchers (e.g., Atwater & Yammrino, 1997; Yammarino & Atwater, 1993) claim that self-other agreement offers additional value to MSF as accurate estimators would have higher performance than over-estimators. This claim has been supported by dozens of studies with U.S. samples (e.g., Atwater et al., 1998; Atwater et al, 2005, Ostroff et al., 2004; Van Velsor, Taylor & Leslie, 1993). An interesting area for research is to find out whether the pattern of relationship persists in different cultures.

Based on what have been found in the Atwater et al. (2005) study and this study, it seems clear that the results achieved from U.S. samples are the exception rather than rule. Self-ratings were not related to performance in the European countries (Atwater et al., 2005) and the GDT sample in this study, and thus self-other agreement provided little additional value in these settings. Self-ratings did relate to performance and accurate estimators with high self-ratings were found to have high performance in the Chinese sample, but the response surfaces displayed a different pattern of relationships. The response surfaces in the Chinese culture were rather flat, with a weak leveling effect at the low end of the line of perfect agreement. Therefore, the results found with the U.S. samples may not be replicable in other cultural settings.

If such a conclusion holds with more studies, the underlying psychological mechanisms that cause self-other agreement effect are more likely to be unique in the U.S. culture. Past research has suggested that such psychological

constructs as self-awareness, self-monitoring, self-esteem, and self-enhancement bias may contribute to the self-other agreement effect (e.g., Fletcher, 1997; London & Smither, 1995; Wohlers & London, 1989). However, these constructs may reflect a unique western view of the individual as an independent, self-contained, and autonomous entity (Markus & Kitayama, 1991). As a result, they may not be as relevant in other cultures. People from a collectivist culture attend to others more than their individualist counterparts (Markus & Kitayama, 1991), so self-awareness may play a less significant role in such a culture. Individualists practice more self-enhancement than collectivists while the latter may instead value self-criticism (Kitayama, Markus, Matsumoto, & Norasakkunkit, 1997), so self-enhancement bias may be of less relevance in a collectivist culture.

However, it should be noted that the conclusion regarding the uniqueness of self-other agreement in the U.S. culture is still tentative. The non-American samples have been either small or heterogeneous and may not be representative. In this study, the Chinese sample included only 428 ratees, compared with 3896 U.S. ratees in the Atwater et al. (2005) study. The quadratic and cross-product terms in the polynomial regression usually account for a very small account of variance in performance, often times only about 1% (e.g., Atwater et al., 1998). A large sample size is needed to make self-other agreement effects more observable. Similar to the non-American samples in the Atwater et al. (2005)

study, the GDT sample was also heterogeneous as there were multiple sub-types.

A Revisit of the Framework

The framework proposed in the present study tries to integrate two perception-based models: Hanges et al.'s (2002) connectionist model of leadership perception and Kenny's (1991) weighted-average model of person perception. The former model places emphasis on self-concept, with an attempt to understand how people perceive the same leadership behavior differently. The latter model stresses behavior, being more interest in how people perceive a personality trait with consensus. These two models, when put together, offer exceptional value to understand the rating process of MSF.

Particularly, the integrated framework proposes that consensus between raters and self-other agreement share the same process. Raters or ratees observe/recall behaviors, process them through self-concepts, and form perceptions. Agreement between any two individuals will be determined by factors that affect behavior and self-concept. Culture is such a factor that can exert its influence through self-concept and collective behavioral patterns.

While some cultural effects were identified in this study, they didn't apply to consensus between raters and self-other agreement in the same way. Consensus between raters was stable across different cultural settings while self-other agreement was culturally sensitive. In addition, demographic effects were consistently observed in self-ratings, and the effects were different across

cultures, indicating that self-ratings were heavily sourced from self-concepts. Self-concept seems to play a larger role in self-other agreement than in consensus between raters. In other words, consensus between raters is based more on behavior than on self-concept, with an assumption that behavior will be less influenced by culture.

The assumption that culture has few effects on behavior is not consistent with what has been proposed in the framework. However, this assumption makes sense if we consider the fact that participants in the Chinese and the GDT samples are mostly from multinational companies. These companies usually are the most developed in terms of organizational structure and organizational culture. Behavior will be constrained by these higher-level factors and becomes more consistent. As such, the speculation that culture has less impact on behavior is warranted, at least in such settings as multinational companies.

Therefore, it seems reasonable to conclude that consensus between other-ratings will be stable across different cultural settings because raters tend to base their ratings on behaviors, while these job-related behaviors are not culturally bound. On the contrary, self-other agreement will vary in different cultural settings because self-ratings are heavily sourced from self-concept in the condition of ambiguity, while self-concept is culturally saturated. This conclusion doesn't necessarily mean that other-ratings are effective in any settings. Other contextual factors, such as organizational structure and organizational culture,

may play a role here, and thus the effectiveness of other-ratings may vary in different organizations.

Limitations

There are several limitations in this study that researchers should be aware of in future research. First, the archival data set used in this study does not have individual ratings. Therefore, all analyses of this study are based on aggregated scores. Using aggregated scores is only an indirect way to investigate culture's effects on self-other agreement and performance. Cultural effects (e.g., the effects of in-group practices) can be either confounded with source effects or canceled out with each other as the result of aggregation. In addition, aggregation may help to improve the reliabilities of subordinate- and peer-ratings. Consequently, correlations between other-ratings will appear more stable in different settings than self-other correlations. Research based on individual scores (e.g., Mount et al., 1998; Scullen, Mount, & Goff, 2000) should help us to better understand cultural effects and is able to provide direct validation to the framework proposed in this study.

Second, the samples used in this study are not totally random, especially for the Chinese and the GDT samples. The participants are managers and executives who have participated in leadership development programs offered by a management consulting firm. Most of them are from big companies and multinational companies, primary users of MSF. Nevertheless, these companies cover a wide array of industries and the participants represent very diverse

functions. The U.S. sample was randomly selected from more than 20,000 managers and executives who worked in the U.S. as non-expatriates. Therefore, the sample should be able to represent the population adequately. Constrained by sample size, convenience sampling was used for the other two samples. Similar to the U.S. sample, these two samples also cover a wide array of functions and industries. In addition, the number of participants from each company was limited to less than 10% of the total participants in a sample, in order to ensure unbiased representativeness. Therefore, the Chinese and the GDT samples are free from dominant influences of a company, a function, or an industry.

Third, GDT is not clearly defined and its sub-types are not specified. This study loosely defined GDT as a unit with members working in different countries, instead of an interdependent work group comprising of culturally diverse members based in two or more nations who share a collective responsibility for making or implementing decisions related to a firm's global strategy (Cohen & Mankin, 1999). As hindsight, this study identified four sub-types of GDT and found the relationship between self-other agreement and performance was different in these sub-types. Future studies should clearly define which sub-type of GDT to examine and collect a large sample accordingly.

Fourth, as this was an exploratory study, moderators were not included. Some influential moderators could include cultural distance (Church, 1982) and management level. Cultural distance may be important in GDTs. For an

American manager, working with a British subordinate is a different assessment context from working with a Japanese subordinate. The relationship of self-other agreement and performance can be different in these two situations too. The moderating effects of job type have been well documented (e.g., Conway & Huffcut, 1997). Correlations between sources are generally lower for managerial jobs than nonmanagerial jobs, due to higher complexity and invisibility in managerial jobs. In GDTs, it is possible that such moderating effects become more prominent due to increasing ambiguities in their work environment. The effects become even more prominent for middle managers and executives as they constantly work across multiple cultural boundaries. Management level may thus become an important moderator.

Fifth, this study used supervisor-ratings of performance, which were collected along with the MSF questionnaire, as the performance index. This may raise some concern for common method errors. While the concern is warranted, the correlation between performance and supervisor-ratings of leadership is at around .75, not an exceptionally high correlation if considering the high reliabilities of these two measures (around .90). Some disparate relationships are also observed with these two measures. For example, the correlations between supervisor-ratings of leadership and other ratings are higher than those between performance and other ratings. Age was not related to supervisor-ratings of leadership but was related to performance. Therefore, the concern for common method errors can be eased.

Finally, related to the sampling issue, more local companies need to be included in future studies. Including the majority of participants from multinational companies may be responsible for weak cultural effects between the U.S. and the Chinese samples. However, it should be noted that more time must pass before Chinese local companies can adopt MSF practices. In addition, language and measurement equivalence issues may become prominent when MSF is used in local companies. From the business perspective, research using participants from multinational companies is still the most valuable.

Future Research and Implications

This study proposes a new framework of self-other agreement that have a great promise to study contextual influences. However, an empirical study based on the U.S., the Chinese, and the GDT samples lends only partial support to the framework. Limitations of the empirical study have been posed and remedies can follow. To isolate cultural effects from other effects, individual scores should be adopted in future studies. In addition, participants from local Chinese companies should be sampled in order to make comparisons between cultures. Furthermore, the moderating effects of cultural distance and management level should be taken into consideration. Finally, more samples from high collectivist cultures, such as Korea and Japan, can be adopted to replicate this study.

This research could be the first attempt to investigate MSF's application in GDTs. Much work has been done to investigate the performance and process of distributed teams (e.g., Cramton, 2001; Hinds & Bailey, 2003; McDonough,

Kahnb, & Barczaka, 2001), but little pays attention to leadership perception in these settings. Although there is limited support for hypotheses concerning GDTs in this study, this line of research should be continued as it might provide critical insights into the MSF process in settings with heterogeneous cultures. Similar suggestions have been proposed to explore feedback-seeking dynamics in multinational companies (Ashford, Blatt, & VandeWalle, 2003). In future studies, sub-types of GDT should be identified. Unique cultural and organizational challenges may exist in different sub-types. When a manager works with his/her subordinates in one nation and his/her peers work in a separate nation, the GDT could be embedded in a hierarchical structure. Consensus between other-ratings will be lower as interaction is usually constrained under such a structure. In contrast, when a manager works with part of his/her subordinates and peers in the same nation, the GDT could be embedded in a team-based structure. Intense interactions among team members may increase consensus between them. Self-other agreement as well as the relationship between self-other agreement and performance can be substantially different in these two situations.

Studies using expatriates as participants could be another promising way to investigate how culture affects the relationship between self-other agreement and performance. Adjustment and performance of expatriates have received a lot of attention from researchers (Black et al., 1991) but little research exists regarding how MSF works for expatriates. Expatriates usually share a different meaning system with his/her subordinates, peers and supervisors, which may

affect self-other agreement. Moreover, expatriates working in the U.S. may have different experience with those working in China, and MSF may work differently too. The framework proposed in this study can help to develop hypotheses for different expatriate samples. Results from these samples will further our understanding of culture's impact on self-other agreement.

A comparison between MSF and assessment centers conducted by this study suggests that the more structured the process the more accurate the rating. This notion receives some support from the GDT sample, where consensus between other-ratings remained high despite cultural, language, and geographical barriers. The structured environment of GDT may help to improve the degree of overlap and reduce unique impressions, which then compensate the unfavorable effects of dissimilar meaning systems and constrained communication. Future studies could include structuredness of the work environment as a moderator.

The results of this study also echo the question raised by Atkins and Wood (2002) regarding what criterion should be chosen for self-other agreement study. Atkins and Wood found self-ratings were negatively related to assessment center results and thus self-other agreement effects were not present with objectively measured criterion. However, their study was based on only 63 participants, making the finding less reliable. Moreover, Ostroff et al. (2004) observed some pronounced self-other agreement effects with compensation as criterion. Therefore, both subjective and objective criteria are appropriate for self-

other agreement studies. Atwater et al. (2005) used one-item competency measure as criterion in their study, which may contribute to some differences found between this study and theirs. It seems that self-other agreement effects are more pronounced with a proximal criterion and become less prominent with a distal criterion. Compensation is a proximate criterion because supervisor-ratings of performance are often used to determine compensation level. Future studies should play close attention to which criterion to use and clearly state the construct linkages between predictor and criterion. More research using different criteria will be welcome.

Another interesting research question is whether consensus between other-ratings and self-other agreement share the same process. Many studies (e.g., Atwater & Yammarino, 1997; Ostroff et al., 2004) have tried to explore how self- and other-ratings are affected by different factors, but little work has been done to clarify the processes. This study attempts to integrate the processes of consensus between other-ratings and self-other agreement into one shared process. However, the results in this study indicate consensus between other-ratings is stable across different settings while self-other agreement is not, suggesting that they may belong to separate processes. Nevertheless, the proposition for one shared process has not been ruled out, since it is possible that self-concept has disparate effects on self- and other-ratings in the same process. Future research should be conducted to answer this question.

Implications for Practice

From the practical standpoint, this study offers some useful guidelines to the feedback section of MSF. A feedback section is an occasion where a coach (a HR personnel or a consultant) goes through the feedback report, which has integrated ratings from different perspectives, and tries to make sense of the information on the report, with the ratee. A feedback section is more a dialog between a coach and a ratee than a mere information exchange process. Under the assistance of the coach, the ratee will gain insights about himself/herself and become motivated to make changes. One key factor of success in this process is to understand self-ratings and their relationship to other-ratings. In the U.S, when both self-ratings and other-ratings are related to performance, self-other agreement provides additional insights. For the European countries, self-ratings may be less relevant and self-other agreement provides little additional information, so emphasis of the feedback section should be placed upon other-ratings. In China, although both self- and other-ratings are related to performance, the relationship is generally additive, which means self-other agreement also adds little value to the feedback section. In GDTs, self-ratings seem to be unrelated to performance but it may result from conflicting meaning systems. Self-ratings should not be treated as errors. Instead, a different epistemological approach can be adopted. Ashford et al. (2003) noted that intercultural communication differences may increase the effort costs of feedback seeking

and cause self-other discrepancies. The feedback section can be seen as an opportunity to facilitate such intercultural communication.

This study has important implications for MSF practice in multinational companies. First, this study indicates that culture has little impact on other-ratings, so multinational companies can readily use MSF for leadership development and other purposes. Subordinates, peers, and supervisors provide some unique perspectives about the ratees, while their ratings are all inter-correlated. This message is especially important to GDTs, where team members are located in different nations. Ratings from people who work distantly are equally informative. Second, this study offers some tips that can be used in MSF data interpretation. For example, subordinate-ratings tend to provide more information related to performance than peer-ratings in China, so peer-ratings should not be over-emphasized. In GDTs, peer-ratings, stead of subordinate-ratings, provide more performance-related information, and thus should be stressed. These tips may help to facilitate a dialog between a coach and a ratee, but should not be overstated.

Continuous improvement on MSF process is also expected in practice. Improvements can include better rating format, more effective rater training, and even different ways to conduct MSF. As an example, the ratee can be asked to come up with a list of critical incidents that both he/she and raters have been involved in. These critical incidents can be seen as work samples or simulations in assessment centers. Raters can then recall the ratee's behaviors in these

incidents and discuss their effectiveness. By imposing more structure, self-other agreement and consensus between other-ratings can be greatly improved.

CHAPTER SIX

SUMMARY

This study intends to follow up a previous study conducted by Atwater et al. (2005), which compared the relationship between self-other agreement and performance across several countries, including the U.S. and five other European countries, with three different samples. In addition to a Chinese sample, which has been expected by Ostroff et al, this study also includes a sample composed of globally dispersed teams (GDTs). Results from these additional samples are valuable for fully understanding the relationship between self-other agreement and performance as well as culture's role in the relationship.

A thorough review of the history and current states of MSF and self-other agreement studies indicates a need to go beyond the narrow psychometric perspective and reach out to other research fields. A comprehensive comparison between MSF and assessment centers suggests MSF should take contextual factors into consideration and structure the process as much as possible. Leadership perception theories (e.g., Hanges et al, 2000) and person perception theories (e.g., Kenny, 1991) provide additional insights into the rating process of MSF, with the former emphasizing self-concept and the latter focusing on behavior and consensus. Two current models of self-other agreement (Campbell, 1985, Yammarino & Atwater, 1993) were then introduced, with their strengths and weaknesses highlighted. A new framework was proposed based on these

pre-existing models and related theories from leadership and person perception fields.

The new framework delineates how two raters observe one's behaviors, process them through self-concepts, and reach ratings under the influence of a wide array of contextual factors. It provides a dynamic and multi-level account of the consensus/self-other agreement process. In this framework, six factors proposed by Kenny (1991): acquaintance, overlap, behavioral consistency, similar meaning systems, communication and unique impression, play an important role in consensus making and self-other agreement. The first three factors are specifically related to behaviors, but this framework doesn't demand all ratings should be based on behavior. Self-concept will be the main source of self-ratings when a behavior has been repeated or observed numerous times, inconsistent behaviors are present, or no behavior can be readily recalled. While the alternative approach is efficient, it also results in underestimation or overestimation. Under this framework, culture plays its role in two different ways. First, it affects the self-concept or the meaning system. Second, it affects collective behavior patterns people interact, such as how to display behaviors in different situations and how to communicate with each other.

The Chinese culture was depicted as a culture characterized with high collectivism/high power distance, which is the opposite of the U.S. culture. Being in a high collectivist/high power distance culture, Chinese are concerned about other's perceptions than Americans. In addition, Chinese pay more attention to

context and are inclined to interpret behavior in the context, and thus rely less on self-concept to derive ratings. Accordingly, it was hypothesized that self-other correlations in the Chinese culture would be higher than that in the U.S. culture. The hypothesis is not supported. While Chinese may attend to other's expectations more, they have the same tendency to base their ratings on self-concept when ambiguities arise. Nevertheless, they may call up the collective self-concept, as evidenced with strong demographic effects found in Chinese ratees and their supervisors. Chinese workplace is believed to be more fragmented than that in the U.S., due to in-group practices and status differences. However, the notion that in-group practices and status differences will result in lower correlations between other-ratings is not supported. Although the correlation between peer- and supervisor-ratings in the Chinese culture is lower than that in the U.S. culture, in-group practices and status differences seem to play limited part in it. It was hypothesized that overestimation would be particularly problematic in China as failing to understand other's expectations would be especially detrimental. However, the overestimation effect was observed only in some extreme conditions where self-ratings were very high and other-ratings were very low. Underestimation may reflect such Confucian virtues as humility and empathy, which can help a ratee effectively respond to other's expectations and gain others' support, and thus achieve acceptable results. The results in this study lend general support to this notion. Consistent with the results in the U.S. sample, accurate estimators with high self- and other-ratings

tended to have high performance in the Chinese sample. However, an unexpected leveling effect was present at the lower end of the line of perfect agreement in the Chinese sample, which may suggest that Chinese raters are less capable of differentiating low performers.

This study also investigated the relationship between self-other agreement and performance in GDTs. According to the framework proposed in this study, the existence of multiple different meaning systems makes meeting other's expectations particularly difficult. Failing to meet other's expectations could result in serious performance issues. In face of more ambiguities, people may also rely more on self-concept to derive ratings. Therefore, it was hypothesized that self-other correlations would be lower in the GDT sample than the U.S. sample. While noticeable differences were observed, they failed to reach the significance level. The existence of multiple different meaning systems may also impede raters to reach consensus, but the use of communication technologies may actually make the situation more structured, and thus help to improve consensus between ratings. Results from this study lend some support to this latter notion. In GDTs, overestimation may reflect a ratee's tendency to be insensitive to cultural differences or overly judgmental, which may result in poor performance. Underestimation may indicate lack of self-confidence in multi-cultural settings as well as poor adjustment to different cultures, and thus will also cause poor performance. However, it was found that self-ratings were not particularly relevant to performance in GDTs, due to weak effects of self-ratings on

performance. Therefore, the proposition that both overestimation and underestimation will cause poor performance is not supported. However, a follow-up analysis suggests that weak effects of self-ratings on performance may only show up in situations where there are more than two different meaning systems in the team.

Several limitations of this study were proposed. First, only aggregated scores were used and cultural effects might have be cancelled out or confounded with other effects. Second, this study was based on archival data. Although efforts were taken to minimize sampling errors, the samples might not be well representative. Third, GDT should be clearly defined and sub-types of GDT should be specified. Fourth, moderators, such as management level and cultural distance were not included in the study. Finally, participants in the Chinese sample were mainly from multinational companies and they might not be the best representatives of Chinese culture.

Future studies should use individual scores, more targeted samples, and carefully selected criteria to validate the framework. Studies with other cultures, GDTs and expatriates will also be beneficial. More research on consensus or self-agreement process is highly expected. From the practical perspective, this study indicates that MSF can be applied to multinational companies in different settings. Self-ratings in GDTs may not provide information as valuable as in the U.S., but they can be used to initiate a useful dialog with a ratee and help to understand how the ratee responds to different cultures. In the feedback section,

peer-ratings should be stressed in GDTs but in China subordinate-ratings should be emphasized. The process of MSF can also be improved by asking the ratee to provide a list of critical incidents, based on which raters could achieve higher consensus and more accurate ratings than the traditional MSF.

Overall, this study contributes to the literature in multiple ways. First, a new framework of self-other agreement was proposed by integrating research from multiple fields. This framework can be used to guide future self-other agreement studies across a wide array of settings. Second, an empirical study was conducted which provided some initial validation to the framework. Some evidence of support has been found, but more future research can be conducted. Third, some important insights have been gained. Specifically, this study indicated that: 1) consensus between raters is stable across culture as other-ratings are based on behavior, which are less influenced by culture; 2) self-other agreement varies in different cultures as self-ratings are strongly influenced by self-concept, and thus are culturally biased; 3) the relationship between self-other agreement and performance typically found in the U.S. could be the exception rather than rule. Finally, implications for practice have been suggested.

REFERENCES

Alimo-Metcalfe, B. (1998). 360 degree feedback and leadership development. *International Journal of Selection and Assessment, 6,* 35-44.

Arthur, W., Jr., & Bennett, W., Jr. (1995). The international assignee: The relative importance of factors perceived to contribute to success. *Personnel Psychology, 48,* 99-114.

Ashford, S. J. (1989). Self-assessments in organizations: A literature review and integrative model. *Research in Organizational Behavior, 11,* 133-174.

Ashford, S. J., Blatt, R., & VandeWalle, D. (2003). Reflections on the looking glass: A reveiw of research on feedback-seeking behavior in organizations. *Journal of Management, 96,* 773-799.

Atkins, P. W. B., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55,* 871-904.

Atwater L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: *Does it really matter? Personnel Psychology, 51,* 577-598.

Atwater, L. E., Waldman, D., Ostroff, C., Robie, C., & Johnson, K. M. (2005). Self-other agreement: Comparing its relationship with performance in the U.S. and Europe. *International Journal of Selection and Assessment, 13,* 25-40.

Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology, 45,* 141-164.

Atwater L. E., & Yammarino, F. J. (1997). Self-other rating agreement: A review and model. In G. R. Frris et al. Eds.), *Research in personnel and human resources management* (pp. 121-174). Greenwich, CT: Jai Press, Inc.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77,* 836-874.

Argyris, C., & Schon, D. A. (1978). *Organizational learning: A theory of action perspective.* Reading, MA: Addison-Wesley.

Arvery, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology, 49,* 141-168.

Axelrod, R. (1984). *Evolution of cooperation.* New York: Basic Books

Antonioni, D. (1994). Improve the performance management process before discontinuing performance appraisals. *Compensation & Benefits Review,26,* 29-37.

Antonioni, D. (1996). Designing an effective 360-degree appraisal feedback process. *Organizational Dynamics, 25,* 24-38.

Baba, M. L., Gluesing, J., Ratner, H, & Wagner, K. H. (2004). The contexts of knowing: Natural history of a globally distributed team. *Journal of Organizational Behavior, 25,* 547-587.

Bandura, A. (1999). Social cognitive theory of personality. In D. Cervone and Y. Shoda (Eds.), *The coherence of personality: Social-cognitive bases of consistency, variability, and organization (pp. 185-241).* New York: The Guilford Press.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38,* 335-345.

Barnes-Farrell, J. L. (2001). Performance appraisal: Person perception processes and challenges. In M. London (Ed), *How people evaluate others in organizations* (pp. 135-153). Mahwh, NJ: Lawrence Erlbaum Associates.

Barnes-Farrell, J. L., & Couture, K. A. (1984). *Effects of appraisal salience on immediate and memory-based judgments* (Tech. Rep. No. 84-1). Honolulu: University of Hawaii, Department of Psychology.

Barrett, R. S. (1966). *Performance rating.* Chicago: Science Research Associates.

Bass, B. M., & Yammarino, F. J. (1991). Congruence of self and others ratings of naval officers for understanding successful performance. *Applied Psychology-An International Review, 40,* 437-454.

Bassett, G. A., & Meyer, H. H. (1968). Performance appraisal based on self-review. *Personnel Psychology, 21,* 421-430.

Bateson, G. (1972). *Steps to an ecology of mind.* New York: Ballantine.

Bernardin, H. J., & Villanova, P. (1986). Performance appraisal. In E. A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 43-62). Lexington, MA: Lexington.

Bernardin, H. J., Dahmus, S. A., & Redmon, G. (1993). Attitudes of first-line supervisors toward subordinate appraisals. *Human Resource Management, 32,* 315-324.

Bernieri, F. J., Zuckerman, M., Koestner, R., & Rosenthal, R. (1994). Measuring person perception accuracy: Another look at self-other agreement. *Personality and Social Psychology Bulletin, 20,* 367-378.

Bingman, W. V. (1939). Halo, invalid and valid. *Journal of Applied Psychology, 23,* 221-228.

Black, J. S., & Mendenhall, M. (1990). Cross-cultural training effectiveness: A review and theoretical framework for future research. *Academy of Management Review, 15,* 113-136.

Black, J. S., Mendenhall, M., & Oddou, G. (1991). Toward a comprehensive model of international adjustment: A integration of multiple theoretical perspectives. *Academy of Management Review, 16,* 291-317.

Bochner, S., & Hesketh, B. (1994). Power distance, individualism/collectivism, and job-related attitudes in a culturally diverse work group. *Journal of Cross-Cultural Psychology, 25,* 233-257.

Borman, W. C. (1997). 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource ManagementReview, 7,* 299-315.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmidt and W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71-98). San Francisco: Jossey-Bass.

Bracken, D. W., Timmreck, C. W., & Church, A. H. (2001a). Preface. In Backen R. W. et al. (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. XXI-XXXIII), San Francisco: Jossey-Bass.

Bracken, D. W., Timmreck, C. W., & Church, A. H. (Eds.). (2001b). *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes.* San Francisco: Jossey-Bass.

Bretz, Jr., R. D., Milkovich, G. T., & Reed, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18,* 321-352.

Brown, T. C. (n. d.). The importance of performance appraisal in North American: A literature review.

Campbell, D. J. (1985). Self-appraisals from two perspectives: Esteem versus consistency influences. *Proceedings of the Midwest Academy of Management,* 110-114.

Campbell, D. (2001). Foreword. In Backen R. W. et al. (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. XIII-XX), San Francisco: Jossey-Bass.

Campbell, D. J., & Lee, C. (1988). Self-appraisal in performance evaluation: Development versus evaluation. *Academy of Management Journal, 13,* 302-314.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III, & Weick, K. E., Jr. (1970). *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill.

Chen, C. C., Chen, X. P., & Meindl, J. R. (1998). How can cooperation be fostered? The culture effects of individualism-collectivism. *Journal of Management Review, 23,* 285-304.

Chinese Culture Connection. (1987). Chinese values and the search for culture-free dimensions of culture. *Journal of Cross-Cultural Psychology, 18,*143-164,

Church, A. T. (1982). Sojourner adjustment. *Psychological Bulletin, 9,* 540-572.

Church, A. H. (1994). Managerial self-awareness in high-performing individuals in organizations. *Journal of Applied Psychology, 82,* 281-292.

Church, A. H. (1997). Do you see what I see? An exploration of congruence in ratings from multiple perspectives. *Journal of Applied Psychology, 27,* 983-1020.

Church, A. H., & Bracken, D. W. (Eds.). (1997). 360-degree feedback systems [Special Issue]. *Group & Organization Management, 22* (2).

Church, A. H., Waclawski, J., & Burke, W. W. (2001). Multisource feedback for organization development and change. In Backen R. W. et al. (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 301-317), San Francisco: Jossey-Bass.

Cohen, S., & Mankin, D. (1999). Collaboration in the virtual organization. *Trends in Organizational Behavior, 6,* 105-120.

Coates, D. (1998). Don't' tie 360 feedback to pay. *Training, 35,* 68-75.

Coens, T., & Jenkins, M. (2000). *Abolishing performance appraisals: Why they backfire and what to do instead.* San Francisco, CA: Berrett-Koehler Publishers.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68,* 1152-1162.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331-360.

Cousins, S. D. (1989). Culture and selfhood in Japan and the United States. *Journal of Personality and Social Psychology, 56,* 124-131.

Costa, P. T, & McCrae, R. R. (1985). *The NEO personality inventory manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Cramton, C. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science, 12,* 346-371.

Day, D., Schleicher, D.J., Unckless, A.L., & Hiller, N.J. (2002). Self-monitoring personality at work: A meta-analytic investigation of construct validity. *Journal of Applied Psychology, 87,* 390-401.

Deming, E. (1986). *Out of the crisis.* Cambridge, MA: Massachusetts Institute of Technology Center for Advanced Engineering.

DeNisi, A. S. (1996). *Cognitive approach to performance appraisal: A program of research.* London, UK: Routledge.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33,* 360-396.

DiNisi, A. S., & Summers, T. P. (1986). *Rating forms and the organization of information: A cognitive role for appraisal instruments.* Paper presented at the 46th Annual Meeting of the Academy of Management. New Orleans, LA.

Drucker, P. (1993). *Post-capitalist society*. New York: Harper Collins.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for heath, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69-102.

Earley, P. C. (1993). East meets west meets mideast: Further explorations of collectivistic and individualistic work groups. *Academy of Management Journal, 36,* 319-348.

Earley. P. C, & Gibson, C. B. (1998). Taking stock in our progress on individualism-collectivism: 100 years of solidarity and community. *Journal of Management, 24,* 285-304.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes, 58*, 51-100.

Edwards, J. R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes, 59*, 307-324.

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36,* 1577-1613.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37,* 1097-1126.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist. 35,* 790-806.

Fahr, J. L., Dobbins, G., & Cheung, B. S. (1991) Cultural relativity in action: A comparison of self-ratings made by Chinese and US workers. *Personnel Psychology,* 44, 129-147.

Farh, J. L., Tsui, A. S., Xin, K. R., & Cheng, B. S. (1998). The influence of relational demography and guanxi: The Chinese case. *Organization Science, 9,* 471-487.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127-148.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117-140.

Fiske, S. T. (1993) Social cognition and social perception. *Annual Review of Sociology, 44,* 155-194.

Fiske, D., & Cox, J. A. (1960). The consistency of ratings by peers. *Journal of Applied Psychology, 44,* 11-17.

Fletcher, C. (1997). Self-awareness-a neglected attribute in selection and assessment? *International Journal of Selection and Assessment, 5,* 183-187.

Fletcher, C., & Baldry, C. (2000). A study of individual differences and self-awareness in the context of multi-source feedback. *Journal of Occupational and Organizational Psychology, 73,* 303-319.

Fox, S., & Dinur, Y. (1988). Validity of self-assessment: A field evaluation. *Personnel Psychology, 41,* 581-592.

Friedman, T. L. (2005). *The World is Flat: A Brief History of the 21st Century.* New York: Farrar, Strauss and Giroux.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75-90.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652-670.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69,* 656-672.

Furnham, A., & Stringfield, P. (1994). Congruence of self and subordinate ratings of managerial practices as a correlate of supervisor evaluation. *Journal of Occupational and Organizational Psychology, 67,* 57-67.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1985). *Metaanalyses of assessment center validity.* Paper presented at American Psychological Association Convention.

Gilbreth, F. B. (1909). *Bricklaying system.* Chicago: Myron C. Clark.

Gilbreth, F. B. (1911). *Motion study.* New York: Van Nostrand.

Gillespie, T. L. (2005). Internationalizing 360-degree feedback: Are subordinate ratings comparable? *Journal of Business and Psychology, 19,* 361-282.

Goodstein, L., Stringfield, P., & Zevulun, E. (1991). *Differences in self and other ratings between Chinese and Caucasian managers.* Paper presented at the Applied Psychology Conference, Tokyo.

Gordon, L.V., & Medland, F. F. (1965). The cross group stability of peer ratings of leadership potential. *Personnel Psychology, 18,* 173-177.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 122,* 373-382.

Greguras, G. J., Robie, C., & Born, M. P. (2001). Applying the social relations model to self and peer evaluations. *Journal of Management Development, 20,* 508-525.

Hall, D. T. (1983). The effect of the individual on an organization's structure, style, and process. In F. Landy, S. Zedeck & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 11-30). New Jersey: Lawrence Erlbaum Associates Publishers.

Handfield-Jones, H. (2000). How executives grow. *The McKinsey Quarterly, 1,* 117-123.

Hanges, P. J., Lord, R. G., & Dickson, M. W. (2000). An information processing perspective on leadership and culture: A case for connectionist architecture. *Applied Psychology: An International Review, 49,* 133-161.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43-62.

Haslam A. A., Powell, C., & Turner, J. C. (2000). Social identity, self-categorization, and work motivation: Rethinking the contribution of the group

to positive and sustainable organizational outcomes. *Applied Psychology: International Review, 49,* 319-339.

Hedge, J. W., Borman, W. C., & Birkeland, S. A. (2001). History and development of multisource feedback as a methodology. In Backen R. W. et al. (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 15-32). San Francisco, CA: Jossey-Bass.

Heine, S. J., Kitayama, S., Lehman, Dr. R., Takata, T., Ide, E., Leung, C., & Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America: An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology, 81,* 599-615.

Heneman, H. G., III. (1974). Comparisons of self- and superior ratings of managerial performance. *Journal of Applied Psychology, 59,* 638-642.

Heneman, R. L., Greenberger, D. B., & Annonyuo, C. (1989). Attributions and exchange: The effects of interpersonal factors on the diagnosis of employee performance. *Academy of Management Journal, 32,* 466-476.

Heneman, R. L., & Wexley, K. N. (1983). The effects of time delay in rating and amount of information observed on performance rating accuracy. *Academy of Management Journal, 26,* 677-686.

Heneman, R. L., Wexley, K. N., & Moore, M. L. (1987). Performance-rating accuracy: A critical review. *Journal of Business Research, 15,* 431-448.

Hezlett, S.A., Ronnkvist, A.M., Holt, K.E. and Hazucha, J.F. (1997) *The PROFILOR^{R} technical summary.* Minneapolis, MN: Personnel Decisions International.

Hinds, P. J., & Bailey, D. E. (2003). Out of sight, out of sync: Understanding conflict in distributed teams. *Organizaiton Science, 14,* 615-632

Ho, D. Y. F. (1976). On the concept of face. *American Journal of Sociology, 81,* 867-884.

Hofstede, G. (1980a). *Cultural consequences*: *International differences in work-related values*. Bevery Hills: Sage Publication.

Hofstede, G. (980b). Motivation, leadership and organization: Do American leadership theories apply abroad? *Organizational Dynamics, 9,* 42-63.

Hofstede, G. (1998). Attitudes, values, and organizational culture: Disentangling the concepts. *Organization Studies, 19,* 477-492

Hofstede, G. & Bond, M. (1988). Confucius and economic growth: New trends in culture's consequences. *Organizational Dynamics, 16,* 4-21.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior, 17,* 627-637.

Hogan, R., Shelton, D. (1998). A socioanalytic perspective on job performance. *Human Performance, 11,* 129-144.

Hollander, E. P. (1954). Buddy ratings: Military research & industrial applications. *Personnel Psychology, 7,* 385-393.

Hollander, E. P. (1956). The friendship factor in peer nominations. *Personnel Psychology, 9,* 435-447.

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology, 41,* 85-90.

Hunter, J. E,, Schmidt, E L., & Jackson, G. B. (1982). *Meta-analysis: Accumulating research findings across studies.* Beverly Hill, CA: Sage.

Ilgen, R. R., & Feldman, J. M. (1983). Performance appraisal: A process approach. In L. L. Cummings and B. M. Staw (Eds.), *Rsearch in organization behavior* (Vol. 5, pp. 141-197). Greenwich, CT: JAI Press.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational behavior and Human Decision Processes, 54,* 321-368.

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review, 10,* 311-321.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219-229.

James, L., & Jones, A. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin, 18,* 1096-1112.

Jones, A. P., & James, L. R. (1979). Psychological climate : Dimensions and relationships of individual and aggregated work environment perceptions. *Organizational Behavior and Human Performance, 23,* 201-250.

Jung, D. I., & Avolio, B. J. (1999). Effects of leadership style and followers' cultural orientation on performance in group and individual task conditions. *Academy of Management Journal*, *42*, 208-218.

Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations* (2nd ed.). New York: Wiley.

Keller, R. T. (1992). Transformational leadership and the performance of research and development project groups. *Journal of Management, 18,* 489-501.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98,* 155-163.

Kenny, D. A. (1994). *Interpersonal perception.* New York: Guilford Press.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102,* 390-402.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional, forced-choice performance evaluation scale. *Journal of Applied Psychology, 65,* 507-516.

Kirkman, B. L., Lowe, K. B., & Gibson, C. B. (2006). A quarter century of Culture's Consequences: a review of empirical research incorporating

Hofstede's cultural values framework. *Journal of International Business Studies, 37,* 285-320.

Kitayama, S. (2002). Culture and basic psychological processes-toward a system view of culture: Comment on Oyserman et al. (2002). *Psychological Bulletin, 128,* 89-96.

Kitayama, S. & Markus, H. R. (1999). Yin and Yang of the Japanese self: The cultural psychology of personality coherence. In D. Cervone and Y. Shoda (Eds.), *The coherence of personality: Social-cognitive bases of consistency, variability, and organization* (pp. 242-302). New York: The Guilford Press.

Kitayama, S. & Markus, H. R., Matsumoto, H., & Norasakkunkit, V. (1997). Individual and collective processes in the construction of the self: Self-enhancement in the United States and self-criticism in Japan. *Journal of Personality and Social Psychology, 72,* 1245-1267.

Klimoski, R. J., & Donahue, L. M. (2001). Person perceptions in organizations: An overview of the field. In M. London (Ed), *How people evaluate others in organizations* (pp. 5-44). Mahwh, NJ: Lawrence Erlbaum Associates.

Knight, F. B. (1923). The effect of the "acquaintance factor" upon personal judgments. *Journal of Educational Psychology, 14,* 129-142.

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin, 106,* 395-409.

Kwan, V. S. Y., John, O. P., Kenny, D. A., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review, 111,* 94-110.

Lance, C. E. (1994). Test of a latent structure of performance ratings derived from Wherry's (1952) theory of ratings. *Journal of Management, 20,* 757-771.

Landy, F. J., & Farr, J. L. (1980). Performance rating, *Psychological Bulletin, 87,* 72-107.

Lauterbach, K. E., & Barnes-Farrell, J. L. (1989). *The effect of congruence between encoding and retrieval processes in performance appraisal.* Paper presented at the 1st annual meeting of the American Psychological Society. Alexandria, VA.

Lawler, E. E., III. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology, 51,* 369-381.

Lawler, E. E., III. (1994) Performance management: The next generation. *Compensation & Benefits Review, 26,* 16-19.

Leslie, J. B., & Fleenor, J.W. (1998). *Feedback to managers: A review and comparison of multi-rater instruments for management development.* Greensboro, NC: Center for Creative Leadership.

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30,* 881-905.

Lievens, F. (n.d.). Research on selection in an international context: Current status and future directions.

Lievens, F., & Harris, M.M., Van Keer, E., & Bisqueret, C. (2003). Predicting crosscultural training performance: The validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. *Journal of Applied Psychology, 88*, 476-489.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and Signal Detection Theory. *Journal of Applied Psychology, 70,* 66-71.

Lord, R. G., & Brown, D. J. (2004). *Leadership processes and follower self-identity.* Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Lord, R. G., Brown, D. J., & Freiberg, S. J. (1999). Understanding the dynamics of leadership: The role of follower self-concepts in the leader/follower relationship. *Organizational Behavior and Human Decision Processes, 78,* 167-203.

Lord, R. G., Brown, D. J., Harvey, J. L., & Hall, R. J. (2001). Contextual constraints on prototype generation and their multi-level consequences for leadership perception. *The Leadership Quarterly, 12,* 311-338.

Lord, R. G., Foti, R. J., & DeVader, C. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance, 34,* 343-378.

Lord, R. G., & Smith, W. G. (1999). Leadership and the changing nature of work performance. In: D. R. Ilgen, & E. D. Pulakos (Eds.), *The changing nature of performance: implications for staffing, motivation, and development* (pp. 192-239). San Francisco: Jossey-Bass.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability—a review and meta-analysis. *Journal of Applied Psychology, 67,* 280-296.

Maloney, P. W., & Hinrichs, J. R. (1959). A new tool for supervisory self-development. *Personnel, 36,* 46-53.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224-253.

Markus, H., & Wurf, E. (1987). The dynamic self-concept: a social psychological perspective. *Annual Review of Psychology, 38,* 299-338.

Maurer, T. J., & Tarulli, B. A. (1996). Acceptance of peer/upward performance appraisal systems: Role of work context factors and beliefs about managers' development capability. *Human Resource Management, 35,* 217-241.

McCaulley, C. and Lombardo, M. (1990) Benchmarks: An instrument for diagnosing managerial strengths and weaknesses. In K.D. Clark and M.S. Clark (Eds.), *Measures of leadership* (pp. 535-545). West Orange, NJ: Leadership Library of America.

McDonough, E. F., III, Kahnb, K. B., & Barczaka, G. (2001). An investigation of the use of global, virtual, and colocated new product development teams *Journal of Product Innovation Management 18*, 110–120.

McGregor, D. (1957). An uneasy look at performance appraisal. *Harvard Business Review*, *35,* 89-94.

Meek, V. L. (1988). Organizational culture: Origins and weaknesses. *Organization Studies, 9,* 453-473.

Mendenhall, M., & Oddou, G. (1985). The dimensions of expatriate acculturation: A review. *Academy of Management Review, 10,* 39-47.

Meyer, H. H. (1991). A solution to the performance appraisal feedback enigma. *Academy of Management Review, 5,* 68-77.

Miller, J. G. (1984). Culture and the development of everyday social explanation. *Journal of Personality and Social Psychology, 46*, 961-978.

Mitchell, T. R. (1983). The effects of social, task, and situational factors on motivation, performance, and appraisal. In F. Landy, S. Zedeck & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 39-59). New Jersey: Lawrence Erlbaum Associates Publishers

Moore, C. C. (1937). Note on the accuracy of judgment in pupil ability ratings. *Journal of Educational Psychology, 28,* 391-393.

Moses, J., Hollenbeck, G. P., & Sorcher, M. (1993). Other people's expectations. *Human Resource Management, 32*, 283-297

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology, 51,* 557-576.

Mount, M. K., & Scullen, S. E. (2001). Multiource feedback ratings: What do they really measure? In M. London (Ed), *How people evaluate others in organizations* (pp. 155-180). Mahwh, NJ: Lawrence Erlbaum Associates.

Mount, M. K., & Thompson, D. E. (1987). Cognitive categorization and quality of performance ratings*. Journal of Applied Psychology, 72,* 240-246.

Murphy K. J., & Cleveland J. N. (1995). *Understanding performance appraisal: Social, organizational and goal-based perspectives.* Thousand Oaks, CA: Sage

Murphy, K. J. (1992). Performance measurement and appraisal: Motivating managers to identify and reward performance. In W. J. Bruns, Jr. (Ed), *Performance, measurement, evaluation, and incentives* (pp. 37-62). Boston, MA: Harvard Business School Press.

Murphy, K. R., Cleveland, J. N., & Mohler, C. J. (2001). Reliability, validity, and meaningfulness of multisource ratings. In Backen R. W. et al. (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 130-148), San Francisco: Jossey-Bass.

Napier, N. K., & Latham, G. P. (1986). Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology, 39,* 827-837.

Nilsen, D., & Campbell, D. (1993). Self-observer rating discrepancies: Once an overrater, always an overrater? *Human Resource Management, 32,* 265-281.

Ostroff, C., Atwater, L. E., & Feinberg. B. J. (2004). Understanding self-other agreement : A look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology, 57,* 333-375.

Ostroff, C., & Ilgen, D. R. (1985). *The relationship between cognitive categories of raters and rating accuracy.* Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Ouchi, W. G., & Wilkins, A. L. (1985). Organizational culture. *Annual Review of Sociology, 11,* 457-483.

Palermo, D. S. (1983). Cognition, concepts, and an employee's theory of the world. In F. Landy, S. Zedeck & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 97-115). New Jersey: Lawrence Erlbaum Associates Publishers

Pollack, D. M., & Pollack, L. J. (1996). Using 360^{o} feedback in performance appraisal. *Public Personnel Management, 25,* 507-528.

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance rating: An examination of ratee race, ratee gender, and rater level effects. *Human Performance, 9,* 103-119.

Ralston, D. A., Egri, C. P., Stewart, S., Terpstra, R. H., & Yu, K. (1999). Doing business in the 21st Century with the new generation of Chinese managers: A study of generational shifts in work values in China. *Journal of International Business Studies, 30,* 415-427.

Redding, S. G., *&* Ng, M. (1983). The role of "face" in the organizational perceptions of Chinese managers. *International Studies of Management and Organization, 3,* 92-123.

Reinhardt, C. (1985). The state of performance appraisal: A literature review. *Human Resource Planning, 8,* 105-110.

Roberts, T. A., Nolen-Hoeksema, S. (1994). Gender comparisons in responsiveness to other's evaluations in achievement settings. *Psychology of Women Quarterly, 18,* 221-240.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322-327.

Rupe, J. C. (1951). When workers rate the boss. *Personnel Psychology, 4,* 271-289.

Ryan, A. M., Brutus, S., Greguras, G. J., & Hakel, M. D. (2000). Receptivity to assessment-based feedback for management development. *Journal of Management Development,* 19, 252-276.

Salam, S., Cox, J. F., & Sims, H. P., Jr. (1997). In the eye of the beholder: How leadership relates to 360-degree performance ratings. *Group & Organization Management, 22,* 189-229.

Sanchez, J. I., & De La Torre, P. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. *Journal of Applied Psychology, 81,* 3-10.

Schein, E. H. (1990). Organizational culture. *American Psychologist, 45,* 109-119.

Schein, E. H. (1992). *Organizational culture and leadership* (2nd ed.). San Francisco: Jossey-Bass.

Schermerhorn, J., Hunt J., & Osborn, R. (2004). *Organizational behavior* (9th ed.). New York: John Wiley and Sons.

Scullen, S. E. (1997). When ratings from one source have been averaged, but ratings from another source have not: Problems and solutions. *Journal of Applied Psychology, 82,* 880-888.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85,* 956-970.

Shamir, B., Zakay, E., Breinin, E., & Popper, M. (1998). Correlates of charismatic leader behavior in military units: Subordinates' attitudes, unit characteristics, and superiors' appraisals of leader performance. *Academy of Management Journal*, *41,* 387-409.

Shweder, R. A., & Sullivan, M. (1993). Cultural Psychology: Who needs it? *Annual Review of Psychology, 44,* 497-523.

Smith, R. J. (1983). *Japanese society.* Cambridge: Cambridge University Press.

Smith, P. B., Peterson, M. F., & Misumi, J. (1994). Event management and work team effectiveness in Japan, Britain and USA. *Journal of Occupational and Organizational Psychology*, *67*, 33-43.

Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology, 31,* 853-888.

Steers, R. and Sanchez-Runde, C. (2002): Culture, motivation and work behavior. In M. Gannon and K. Newman (eds.), *The Blackwell handbook of cross-cultural management* (pp. 190-216). Oxford: Blackwell Business Publishers.

Sully de Luque, M. F., & Sommer, S.M. (2000). The impact of culture on feedback-seeking behavior: An integrated model and propositions. *The Academy of Management Review*, *25,* 829-849.

Sulsky, L. M., & Day, D. V. (1994) Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79,* 535-543.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). MA: Allyn and Bacon

Tayeb, M. (1994). Organizations and national culture: Methodology considered. *Organization Studies, 15,* 429-446.

Taylor, E. W. (1947). *Principles of scientific management.* New York: Norton. (Original work published 1911).

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25-29.

Thornton, G. C. (1968). The relationship between supervisory and self-appraisals of executive performance. *Personnel Psychology, 21,* 441-456.

Thornton, G. (1980). Psychometric properties of self-appraisals of job performance. *Personnel Psychology, 33,* 263-271.

Thornton, G. C. (1992). *Assessment Centers in Human Resource Management*. Addison Wesley: Reading MA.

Tornow, W. W. (1993a). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management, 32,* 221-229.

Tornow, W. W. (Ed.). (1993b). 360-degree feedback [Special Issue]. *Human Resource Management, 32* (2-3).

Triandis, H. C. (1989). The self and social behavior in differing cultural contexts. *Psychological Review, 96,* 506-520.

Triandis, H. C. (1994). Theretical and methodological approaches to the study of collectivism and individualism. In U. Kim et al. (Eds.), *Individualism and collectivism: Theory, method, and applications* (pp. 41-51). Thousand Orks: Sage Publications.

Triandis, H. C. (1995). *Individualism and collectivism.* Boulder, CO: Westview Press.

Triandis, H. C. (2000). Dialectics between cultural and cross-cultural psychology. *Asian Journal of Social Psychology, 3,* 185-195.

Van Velsor, E., & Leslie, J. B. (1991). *Feedback to managers: A guide to evaluating multi-rater feedback instruments* (Vol 1). Greensboro, NC: Center for Creative Leadership.

Van Veslor, E., Taylor, S., & Leslie, J. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resources Management, 32,* 249-263

.Viswesvaran, C. (1996). Modeling job performance: Is there a general factor? Paper presented at the 11th annual conference of the Society of Industrial and Organizational Psychology, San Diego, CA.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557-574.

Weitz, J. (1958). Selecting supervisors with peer ratings. *Personnel Psychology, 11,* 283-291.

Wherry, Sr., R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521-551.

Wherry, R. J., & Fryer, D. H. (1949). Buddy ratings: Popularity contest of leadership criterion? *Personnel Psychology, 2,* 147-159.

Wiese, D. S., & Buckley, M. R. (1998). The evolution of the performance appraisal process. *Journal of Management History, 4,* 233-249.

Williams, S. S., & Leavitt, H. J. (1947). Group opinion as a predictor of military leadership. *Journal of Consulting Psychology, 11,* 283-291.

Woehr, D. J.. (1994) Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79,* 525-534.

Woehr, D. J., & Huffcutt, A. I. (1994) Rater training for performance appraisals: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189-205.

Wohlers, A. J., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self awareness. *Personnel Psychology, 42,* 225-261.

Wuthnow, R., & Witten, M. (1988). New directions in the study of culture. *Annual Review of Sociology, 14,* 49-67.

Yammarino, F. J., & Atwater, L. E. (1993). Understanding self-perception accuracy: Implications for human resources management. *Human Resources Management, 32,* 231-247.

Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as others see them? Implications of self-other rating agreement for human resources management. *Organizational Dynamics, 25,* 35-44.

Yammarino, F. J. (2003). Modern data analytic techniques for multisource feedback. *Organizational Research methods, 6,* 6-14.

Yang, K. S. (1981). Social orientation and individual modernity among Chinese students in Taiwan. *Journal of Social Psychology, 113,* 159-170.

Yau, O. H. M. (1986). Chinese cultural values and their marketing implications. In R. T. Hsieh and S. A. Scherllng (Eds.), *Proceedings of the Academy of International Business Southeast Asia Regional Conference* (pp. 215-237). Taipei: National Chiao Tung University.

Yu, A. B., & Yang, K. S. (1994). The nature of achievement motivation in collectivist societies. In U. Kim et al. (Eds.), *Individualism and collectivism: Theory, method, and applications* (pp. 239-250). Thousand Orks: Sage Publications.